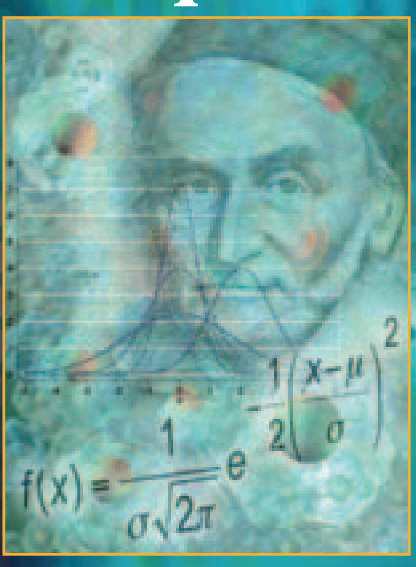
L'innovation thérapeutique en cancérologie **5**

Méthodes biostatistiques appliquées à la recherche clinique en cancérologie



Coordonnateurs
Andrew Kramar
et Simone Mathoulin-Pélissier





Collection - L'innovation thérapeutique en cancérologie 5

Méthodes biostatistiques

appliquées à la recherche clinique en cancérologie

Andrew Kramar
Simone Mathoulin-Pélissier

La biostatistique est indispensable à toute recherche clinique, de la conception de l'étude à l'interprétation des résultats en passant par le calcul de l'effectif et l'analyse statistique.

En cancérologie, le statisticien, confronté à des demandes souvent complexes, doit développer et utiliser des outils de plus en plus sophistiqués. Il devient un véritable détective face à la quantité de données à traiter pour faire émerger les résultats et formuler les hypothèses qui seront discutées avec les investigateurs. Car leur objectif est le même : trouver des outils adaptés pour mieux définir les stratégies thérapeutiques de demain.

Ce nouveau volume de la collection « L'innovation thérapeutique en cancérologie » a été conçu pour faire découvrir aux lecteurs les méthodes biostatistiques les plus couramment utilisées dans les études cliniques interventionnelles ou observationnelles.

Il s'adresse aux biostatisticiens, cliniciens, biologistes, pharmaciens, data-managers, assistants de recherche clinique, étudiants, ainsi qu'aux statisticiens impliqués dans des domaines autres que le cancer.

Rappelant les notions essentielles à la compréhension de la variabilité statistique avant d'aborder la spécificité de la cancérologie, les auteurs guident le lecteur dans le foisonnement des méthodes, donnent les clés pour déterminer la méthodologie adaptée à la réalisation des projets de recherche.

Illustré de nombreux exemples, cet ouvrage permettra au lecteur de mieux situer les exigences statistiques lors de l'élaboration et l'analyse de ses projets de recherche et sera un guide de référence pour les chercheurs travaillant dans tous les domaines de la cancérologie.

Collection réalisée en partenariat avec la SFC





www.jle.com

Méthodes biostatistiques appliquées à la recherche clinique en cancérologie

Une collection dirigée par
Jacques Robert
Professeur de cancérologie biologique
Université Victor-Segalen, Institut Bergonié, Bordeaux

Réalisée en partenariat avec
la Société française du cancer



L'innovation thérapeutique en cancérologie

Méthodes biostatistiques appliquées à la recherche clinique en cancérologie

Coordonnateurs
Andrew Kramar et Simone Mathoulin-Pélissier





ISBN 978-2-7420-0774-5

Éditions John Libbey Eurotext
Éditions John Libbey Eurotext
127, avenue de la République
92120 Montrouge, France
Tél : 01 46 73 06 60
e-mail : contact@jle.com
http://www.jle.com

Éditrice : Maud Thévenin

Auteurs

Bernard Asselain, Service de biostatistique, Inserm U900, Institut Curie, Paris

Anne Aupérin, Unité de méta-analyse, Service de biostatistique et d'épidémiologie, Institut Gustave-Roussy, Villejuif

Agathe Bajard, Unité de biostatistique et d'évaluation des thérapeutiques, Centre Léon-Bérard, Lyon

Caroline Bascoul-Mollevi, Unité de biostatistique, Centre de traitement des données, Centre Val d'Aurelle Paul-Lamarque, Montpellier

Carine Bellera, Unité de recherche et épidémiologie cliniques, Centre de traitement des données et Inserm CIC-EC7, Institut Bergonié, Bordeaux

Ellen Benhamou, Service de biostatistique et d'épidémiologie, Institut Gustave-Roussy, Villejuif

Virginie Berger, Institut de cancérologie de l'Ouest – Paul-Papin, Centre d'évaluation clinique en oncologie, Angers

Camille Berneur-Morisseau, Centre d'évaluation clinique – data management, Institut de cancérologie de l'Ouest – René-Gauducheau, Saint-Herblain

Jean-Marie Boher, Équipe biostatistique, Bureau d'études cliniques, Unité de recherche clinique, Institut Paoli-Calmettes, Marseille

Franck Bonnetain, Unité de biostatistique et d'épidémiologie, EA 4184, Centre Georges-François-Leclerc, Dijon

Loïc Campion, Unité de biostatistique, Institut de cancérologie de l'Ouest – René-Gauducheau, Saint-Herblain

Sylvie Chabaud, Unité de biostatistique et d'évaluation des thérapeutiques, Centre Léon-Bérard, Lyon

Emmanuel Chamorey, Unité d'épidémiologie et biostatistiques, Centre Antoine-Lacassagne, Nice

Adélaïde Doussau, CHU de Bordeaux (Unité de soutien méthodologique à la recherche clinique et épidémiologique), Université de Bordeaux (ISPED), Inserm CIC-EC7, Bordeaux

Benjamin Esterni, Équipe biostatistique, Bureau d'études cliniques, Unité de recherche clinique, Institut Paoli-Calmettes, Marseille

Céline Ferlay, Unité de biostatistique et d'évaluation des thérapeutiques, Centre Léon-Bérard, Lyon

Thomas Filleron, Bureau des essais cliniques, Institut Claudius-Régaud, Toulouse

Jocelyn Gal, Unité d'épidémiologie et biostatistiques, Centre Antoine-Lacassagne, Nice

Mélanie Gauthier, Unité de biostatistique et d'épidémiologie, Centre Georges-Francois-Leclerc, Dijon

Sophie Gourgou-Bourgade, Unité de biostatistique, Centre de traitement des données, CRLC Val d'Aurelle Paul-Lamarque, Montpellier

Éléonore Gravier, Service de biostatistique et Département de transfert, Institut Curie, Inserm U900, École des Mines de Paris, Paris

Catherine Guérin-Charbonnel, Unité de biostatistique, Institut de cancérologie de l'Ouest – René-Gauducheau, Saint-Herblain

Catherine Hill, Service de biostatistique et d'épidémiologie, Institut Gustave-Roussy, Villejuif

Nadine Houédé, Département d'oncologie médicale, Institut Bergonié, Bordeaux

Andrew Kramar, Unité de méthodologie et biostatistique, Centre Oscar-Lambret, Lille

Fabrice Kwiatkowski, Unité de recherche clinique, Centre Jean-Perrin, Clermont-Ferrand

Agnès Laplanche, Service de biostatistique et d'épidémiologie, Institut Gustave-Roussy, Villejuif

Isabelle Le Ray, Unité de biostatistique et d'épidémiologie, EA 4184, Centre Georges-François-Leclerc, Dijon et Centre d'investigation clinique plurithématique Inserm 803, CHU, Dijon

Simone Mathoulin-Pélissier, Unité de recherche et épidémiologie cliniques, Centre de traitement des données et Inserm CIC-EC7, Institut Bergonié, et Université Bordeaux Segalen, Bordeaux

Stefan Michiels, Breast Cancer Translational Research Laboratory JC Heuson, Institut Jules-Bordet, Université libre de Bruxelles, Bruxelles

Emmanuelle Mouret-Fourme, Épidémiologie clinique-DIM, Hôpital René-Huguenin, Institut Curie, Saint-Cloud

Xavier Paoletti, Service de biostatistique, Inserm U900, Institut Curie, Paris

Bruno Pereira, Délégation à la recherche clinique et à l'innovation, Centre hospitalier universitaire, Clermont-Ferrand

Marie-Quitterie Picat, CHU de Bordeaux, Université de Bordeaux (ISPED), et Institut Bergonié (Unité de recherche et d'épidemiologie cliniques), Bordeaux

Jean-Pierre Pignon, Unité de méta-analyse, Service de biostatistique et d'épidémiologie, Institut Gustave-Roussy, Villejuif

Raphaël Porcher, Département de biostatistique et informatique médicale, Hôpital Saint-Louis, Paris

Lise Roca, Unité de biostatistique, Centre de traitement des données, Centre Val d'Aurelle Paul-Lamarque, Montpellier

Pascal Roy, Service de biostatistique des Hospices civils de Lyon et Équipe biostatistique santé de l'UMR 5558 CNRS Université Claude-Bernard Lyon 1

Alexia Savignoni, Service de biostatistique, Inserm U900, Institut Curie, Paris

Anne-Lise Septans, Institut de cancérologie de l'Ouest – Paul-Papin, Centre d'évaluation clinique en oncologie, Angers

Simon Thezenas, Unité de biostatistique, Centre de traitement des données, CRLC Val d'Aurelle Paul-Lamarque, Montpellier

Fabien Valet, Service de biostatistique, Institut Curie, Inserm U900, École des Mines de Paris, Paris

Michel Velten, Service d'épidémiologie et de biostatistique, Centre Paul-Strauss, Strasbourg et Laboratoire d'épidémiologie et de santé publique, EA 3430 et Université de Strasbourg

Préface

La cancérologie est certainement la discipline qui évolue le plus rapidement en ce début de siècle. Le nombre des pistes ouvertes par le concept de ciblage thérapeutique est impressionnant ; même si peu de succès définitifs sont encore enregistrés, plus de 700 essais thérapeutiques sont actuellement ouverts dans le monde. L'évaluation des molécules proposées par les laboratoires pharmaceutiques industriels ou académiques requiert une méthode rigoureuse. Plus que toute autre discipline, la cancérologie moderne repose sur la nécessité d'une médecine fondée sur les faits *(Evidence-based medicine)* ; c'est pour apporter ces données que s'impose la rigueur scientifique. La connaissance de la méthodologie n'est pas intuitive ; elle repose d'abord sur une bonne appréhension des statistiques. Cette science apparaît difficile d'abord à bien des médecins ; c'est peut-être parce qu'ils n'ont pas fait l'effort, au début de leurs études, de chercher à en comprendre tout l'intérêt... Et c'est une fois engagés dans la recherche clinique qu'ils s'aperçoivent de la nécessité d'en comprendre, sinon le formalisme mathématique, du moins le langage et les possibilités.

La Société française du cancer est heureuse d'ajouter à sa collection « L'innovation thérapeutique en cancérologie » ce remarquable ouvrage, travail d'équipe de plusieurs années coordonné par Andrew Kramar et Simone Mathoulin-Pélissier, ouvrage qui sera vite indispensable à tous ceux qui participent aux essais cliniques en cancérologie. Innovation thérapeutique ? Oui, la méthodologie statistique innove en permanence : il suffit de regarder en arrière pour voir le chemin parcouru en vingt ans ! Les phases I tirent maintenant parti des outils bayésiens pour optimiser leur *design* et arriver plus vite aux doses efficaces ; le concept de dose maximale tolérée a dû laisser la place, pour les thérapies ciblées, au concept de dose efficace optimale ; les analyses biologiques d'ADN et d'ARN à haut débit sont maintenant intégrées dans des protocoles thérapeutiques destinés à évaluer leur apport réel ; les essais cliniques sélectionnent les patients à inclure en fonction des caractéristiques de leurs tumeurs : tout cela, et bien d'autres choses, était inimaginable il y a vingt ans... Et, bien sûr, toutes ces innovations paraîtront ringardes à nos successeurs, dans vingt ans ! Mais nous aurons, j'espère, renouvelé l'ouvrage d'ici là par plusieurs éditions « revues et augmentées »...

Je voudrais saluer ici le travail accompli par les coordonnateurs de l'ouvrage ; faut-il le qualifier d'herculéen pour montrer l'énergie dépensée ? Ou de bénédictin pour évoquer le soin tatillon avec lequel ils ont défini les thèmes, distribué les chapitres, sélectionné les auteurs ? Les auteurs ont été choisis en fonction de leur intérêt pour le sujet à traiter et de leur notoriété dans ce sujet. Ils ont tous l'habitude du travail d'équipe avec les cliniciens, comme celle de réfléchir ensemble : de la sorte, ils ont bâti un livre très homogène où le lecteur trouvera sans peine ce dont il a besoin pour conduire sa recherche clinique, quel que soit son niveau de formation aux méthodes

biostatistiques. Je suis certain que, des jeunes internes aux investigateurs chevronnés, chacun y trouvera de quoi améliorer sa pratique, même les méthodologistes de métier ! La véritable formation est celle que l'on acquiert toute sa vie, et nos connaissances ne sont jamais définitives.

Les Éditions John Libbey Eurotext, partenaires réguliers de la Société française du cancer, ont relevé le défi de publier un livre de haut niveau, avec des équations de typographie délicate et des tableaux de composition complexe, sans dessins attrayants ni recettes de cuisine alléchantes ; mais je suis certain que ce livre rendra plus service, aux médecins comme aux malades, que bien d'autres *best-sellers* !

Jacques Robert,
responsable de la formation à la Société française du cancer

Sommaire

Introduction

A. Kramar et S. Mathoulin-Pélissier

En recherche clinique, les méthodes biostatistiques sont utilisées tant pour définir des hypothèses de recherche en collaboration avec les cancérologues que pour résumer l'information fournie par la collecte de données issues d'études cliniques ou observationnelles. Ces méthodes sont utilisées pour l'évaluation d'un nouveau test diagnostique, d'une nouvelle stratégie de traitement (essai thérapeutique), de facteurs de risque ou de facteurs pronostiques. Pour comprendre les méthodes statistiques et leur utilité dans une démarche expérimentale, il est nécessaire de connaître la population à laquelle on s'adresse, les conditions dans lesquelles les résultats ont été obtenus et si l'interprétation des conclusions est correcte.

La statistique est d'abord une science enseignée dans les universités, permettant ainsi d'obtenir des diplômes. La Société française de statistique (www.sfds.asso.fr) regroupe des statisticiens de langue française travaillant dans tous les domaines de la statistique : démographie, économie, physique, agriculture ou médecine. La statistique est d'autant plus facile à comprendre que l'étudiant a une formation mathématique antérieure et qu'il désire s'orienter vers des applications pratiques. Les notions de probabilités sont un atout incontestable, car la statistique est aussi un état d'esprit – certains disent même que c'est une philosophie. Malheureusement, elle est souvent assimilée à tort, par un raccourci, à « des statistiques » ou « des chiffres ».

Les statisticiens impliqués dans le domaine de la cancérologie sont confrontés régulièrement à des questions scientifiques demandant des analyses de plus en plus sophistiquées. Ce livre a ainsi comme objectif principal d'apporter un aperçu des méthodes statistiques les plus couramment utilisées en cancérologie. Si l'exhaustivité n'est pas assurée, c'est en raison de la multitude d'articles plus techniques qui paraissent régulièrement dans les revues statistiques telles que le *Journal of the American Statistical Association*, *Biometrics*, *Statistics in Medicine*, *Biostatistics*, *Drug Information Journal* ou le *Biopharmaceutical Journal* pour n'en mentionner que quelques-unes. Des revues spécialisées en cancérologie, comme le *British Journal of Cancer*, *Journal of Clinical Oncology*, *European Journal of Cancer*, *Cancer Clinical Research* ou la *Revue d'Épidémiologie et Santé Publique*, présentent de temps en temps des notes méthodologiques ou pédagogiques en statistiques.

La confrontation (voire l'incompréhension) de la biostatistique avec la médecine provient en partie du fait que le médecin traite d'abord l'individu (un patient) qu'il a devant lui. La difficulté réside dans le fait qu'il n'est pas évident de comprendre pourquoi deux patients, présentant *a priori* les mêmes caractéristiques, ne vont pas réagir de la même manière au même traitement. Un patient va voir sa maladie régresser, alors que l'autre présentera des effets toxiques sans

guérison. La biostatistique devient alors incontournable pour essayer de mieux appréhender ces phénomènes. Le chercheur biostatisticien s'intéresse à tous les individus (sujets ou patients) et aux informations liées et devient alors un détective devant la quantité croissante d'informations nécessaires à traiter pour essayer de comprendre les résultats et formuler des hypothèses de travail, ces hypothèses devant être dès lors discutées avec les cancérologues dans un dialogue productif pour l'avancée des connaissances.

Ainsi, l'objectif est le même pour le clinicien et le biostatisticien : résoudre les bonnes questions avec des outils adaptés à chaque situation pour mieux cibler et individualiser les stratégies diagnostiques et les traitements de demain. **Ce livre spécialement dédié aux méthodes biostatistiques en cancérologie** présente la biostatistique pour ce qu'elle est, un domaine de recherche propre applicable et incontournable dans tout projet de recherche clinique, qui va de la conception de l'étude et des hypothèses sous-jacentes au calcul du nombre de sujets nécessaires jusqu'à l'analyse statistique et à l'interprétation des résultats. Les premiers chapitres font un bref rappel de notions statistiques de base nécessaires pour décrire les variables observées. Cette terminologie est nécessaire pour le calcul et la compréhension des tests statistiques. Les domaines plus spécifiques sont présentés en deux parties permettant de situer le métier du biostatisticien dans la planification de l'essai et dans l'analyse statistique des résultats. Ces deux parties sont présentes dans chaque protocole d'un essai clinique.

L'objectif de ce livre est de rassembler dans un seul volume des méthodes statistiques très utilisées en recherche clinique en cancérologie. Chaque chapitre a un but précis, mais peut nécessiter des prérequis d'autres chapitres. De nombreux exemples montrent l'intérêt et les limites des méthodes présentées. Il pourra être un guide de référence pour les statisticiens travaillant dans tous les domaines de la cancérologie. La bibliographie comporte des références à des revues scientifiques mais aussi à des ouvrages, des sites Internet pertinents, ainsi qu'à quelques logiciels utiles.

Cet ouvrage est destiné aux biostatisticiens, cliniciens, biologistes, pharmaciens, data-managers, assistants de recherche clinique, étudiants, ainsi qu'aux statisticiens impliqués dans des domaines autres que le cancer.

La première partie concerne les distributions statistiques les plus couramment utilisées en cancérologie. Ce sont les premiers outils nécessaires pour appréhender la variabilité statistique et les statistiques descriptives. Une compréhension des tests d'hypothèses devrait permettre au lecteur de bien intégrer les démarches nécessaires pour planifier une étude clinique et affronter les chapitres suivants avec l'œil d'un investigateur averti. Le choix du bon test, la magie du petit p et une introduction aux statistiques bayésiennes clôturent cette partie. La seconde partie résume différents critères de jugement rencontrés dans les études en cancérologie selon que l'on s'intéresse à la réponse, à la tolérance, à la survie, à la qualité de vie ou aux événements comme la survie sans rechute en tant que critère de substitution à la survie globale. La partie III est consacrée à la stratégie des analyses univariées sur les données de survie, les facteurs pronostiques et prédictifs de la réponse aux traitements, les risques compétitifs et quelques aspects importants liés au suivi et à la surveillance des événements. La partie IV est consacrée aux analyses multivariées qui permettent d'évaluer plusieurs variables simultanément pour expliquer la variabilité d'un critère binaire (régression logistique) ou d'un critère de survie (modèle de Cox). Un aperçu des

techniques propres aux méta-analyses ainsi que la particularité des analyses génomiques clôturent cette partie. La partie V est importante en cancérologie puisqu'elle est dédiée à la planification des essais cliniques de phase I, II et III ; nous avons souhaité aussi consacrer l'un de ses chapitres à la particularité des essais dits de phase 0. Enfin, la dernière partie expose des aspects pratiques nécessaires pour la gestion des données cliniques, pour mettre en œuvre la randomisation des patients dans une étude, et présente les rôles et responsabilités des comités de surveillance des essais, la rédaction d'un plan d'analyse et d'un rapport statistique, ainsi qu'un aperçu des outils spécialisés nécessaires pour entreprendre les analyses statistiques.

Nous souhaitons que l'acquisition de connaissances apportées dans ce livre permette aux lecteurs de mieux situer les exigences statistiques (hypothèses et analyses) lors de l'élaboration de projets de recherche en cancérologie, de l'exploitation des données acquises ainsi que lors de la valorisation du travail scientifique en termes de publication.

Enfin, cet ouvrage n'aurait pu se faire sans la participation de nos nombreux collègues, auteurs et co-auteurs qui, patiemment, ont répondu à nos demandes et attendu nos remarques et la publication finale de ce livre. Nous les remercions sincèrement. Cet ouvrage représente donc aussi leur implication dans le monde académique de la recherche clinique en cancérologie, que ce soit dans des activités de soutien méthodologique ou de recherche en biostatistique.

Partie I
La variabilité statistique

Distributions statistiques

V. Berger, C. Guérin-Charbonnel

Pour analyser des données, il est nécessaire dès l'étape de conception du projet d'identifier le type et la nature de la variable à décrire. Cette identification va déterminer les statistiques à mettre en œuvre pour résumer l'information. Ce chapitre décrit brièvement les deux types de variables le plus souvent utilisées en cancérologie, à savoir les variables quantitatives, qui peuvent être quantifiables par une mesure continue, et les variables qualitatives, qui peuvent être classées en deux ou plusieurs catégories.

Différents types de variables

Les variables peuvent être de différents types : quantitatifs ou qualitatifs. Elles peuvent être de nature objective ou subjective. Les variables quantitatives sont le résultat d'une mesure ; elles représentent une « quantité mesurable », telle que la taille, le poids, la dose reçue. Elles sont toujours accompagnées d'une unité de mesure (année, cm, kg, mg, etc.). Les variables qualitatives ne sont pas le résultat d'une mesure ; elles sont dites « non mesurables » et représentent une « qualité » d'un individu, telle que la couleur des yeux, les antécédents familiaux de cancer, la présence d'une pathologie chronique, etc. Les variables *objectives* sont le plus souvent des mesures chiffrées (poids, taille, etc.), mais peuvent également représenter une caractéristique propre à l'individu observé (sexe, couleur des yeux, etc.). Les variables *subjectives* sont dépendantes de l'examinateur ou du sujet (examen neurologique clinique, qualité de vie, douleur, etc.).

Variables quantitatives : la loi normale

Les variables quantitatives ou continues sont des valeurs mesurables, par exemple le poids, une valeur biologique (nombre de globules blancs, créatinine, etc.) ou la concentration d'un médicament. Par convention, on note X la variable qui nous intéresse, en lui associant toujours une unité (kg, mg, mg/mL, etc.). Dans la population générale, la distribution d'une valeur quantitative d'une donnée physiopathologique suit généralement une loi normale. Cette distribution, décrite de manière indépendante par l'Allemand Gauss (1809) et le Français Laplace (1812), est une loi de probabilité continue, $f(x)$, ayant comme formule :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Avec μ la moyenne et σ l'écart-type. On utilise de manière interchangeable les termes de loi normale ou loi gaussienne. La valeur de la constante π vaut 3,14159... La variance est égale à σ^2, le carré de l'écart-type σ. Cette distribution est donc complètement décrite par deux paramètres : la moyenne et la variance. La moyenne est une mesure de valeur centrale qui reflète l'exactitude de la mesure. La variance est une mesure de dispersion qui reflète la précision de la mesure. Cette loi de probabilité Laplace-Gauss est primordiale en statistique. Non seulement elle permet de décrire la plupart des variables quantitatives continues par deux quantités (la moyenne et l'écart-type), mais la loi normale centrée réduite ($\mu = 0$, $\sigma = 1$) sert de jauge ou standard pour la plupart des tests statistiques.

À retenir

Une loi de probabilité est comprise entre 0 et 1.
La loi normale est la condition d'application de beaucoup de tests statistiques.
Une moyenne est d'autant plus représentative de la distribution que la variance est petite.

On peut représenter graphiquement la loi normale par une courbe symétrique dite en forme de « cloche » de part et d'autre de la moyenne *(figure 1)* avec x en abscisse (axe horizontal) et $y = f(x)$ en ordonnée (axe vertical). En comparant plusieurs lois normales centrées autour de la moyenne « 0 », plus la variance augmente, plus la courbe devient plate, et plus l'estimation de la moyenne devient imprécise. Dans cette situation, l'intervalle de confiance augmente également, ce qui se traduit par plus d'hétérogénéité dans la population *(figure 1)*. Lorsque l'on est en présence de deux populations de patients ayant la même variance pour une variable mesurée, il est légitime de comparer les moyennes, comme par exemple la 2e courbe qui est centrée autour de la valeur 0 ($\mu = 0$, $\sigma = 1$) et la 5e courbe qui est centrée autour de la valeur 2 ($\mu = 2$, $\sigma = 1$). On voit que la 5e courbe est décalée vers la droite de deux unités par rapport à la 2e. Les tests statistiques permettant de comparer deux ou plusieurs populations seront présentés dans le chapitre I.4 « Choix du bon test statistique » (*cf.* page 28).

Lorsque les valeurs des variables sont entières, par exemple l'âge du patient exprimé en années, le nombre d'épisodes de neutropénie ou le stade T de la maladie, ces valeurs sont appelées « **variables quantitatives discontinues** ou discrètes ». La représentation de la répartition se fait la plupart du temps en utilisant un diagramme sous forme d'histogramme. La *figure 2* présente la répartition de l'âge sur un échantillon aléatoire de 500 individus d'un âge moyen de 50 ans et un écart-type de 10, sur lequel est superposée la loi normale de même moyenne et variance. Cet exemple montre l'utilité de cet exercice, car les mesures prises dans cet échantillon servent à estimer la distribution de l'âge dans la population générale.

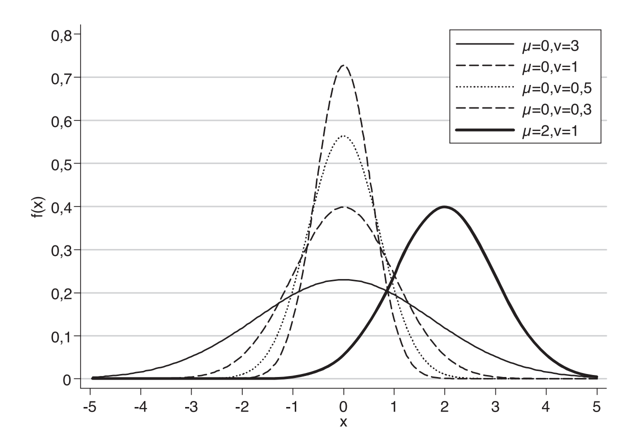


Figure 1. Cinq exemples de lois gaussiennes.

Estimations des paramètres de la loi normale

En statistique, on commence toujours par définir les quantités dont on a besoin. Soit X_i la variable qui nous intéresse, par exemple l'âge du patient. Sur notre échantillon de n patients, on va recueillir la donnée et la représenter par la valeur x_i pour le patient i. Cet indice i varie donc entre 1 et n, 1 pour le premier patient et n pour le dernier. Dans un premier temps, l'ordre des valeurs n'a pas d'importance. On verra par la suite que certains tests statistiques font appel à des données ordonnées.

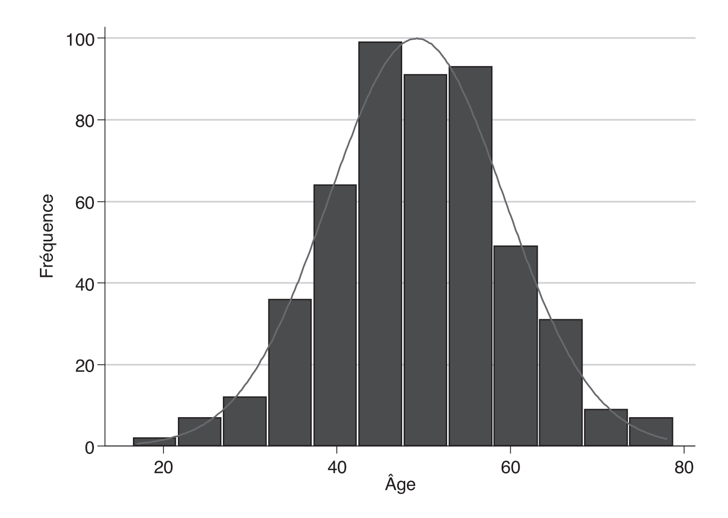


Figure 2. Exemple d'un histogramme.

• La **moyenne** estimée *(m)*, estimation du paramètre μ, est obtenue par la somme de toutes les valeurs observées x_i (de i = 1 à n) divisée par le nombre total d'observations :

$$m = \sum_{i=1}^{n} x_i / n.$$

L'estimation de la moyenne dépend donc de la taille de l'échantillon et des valeurs des observations. En effet, si le dénominateur n est relativement petit, une seule valeur aberrante dans les observations modifiera d'une manière importante la moyenne. Pour l'exemple de la *figure 2*, la moyenne estimée *m* est égale à 49,19.

• La **variance** estimée s^2, estimation du paramètre σ^2, représente la dispersion des valeurs autour de la moyenne dans l'échantillon. Elle est obtenue par la somme des écarts des observations autour de la moyenne estimée divisée par le nombre d'observations moins un :

$$s^2 = \sum_{i=1}^{n} (x_i - m)^2 /(n-1).$$

On peut voir que plus les valeurs x_i sont proches de la moyenne *m*, plus la variance sera petite et vice versa. Dans le cas où tous les sujets ont le même âge, cette variance est égale à zéro et il n'y a donc aucune variabilité. Pour l'exemple de la *figure 2*, la variance estimée s^2 est égale à 103,05.

D'autres estimateurs

Outre la moyenne et la variance, on peut présenter les paramètres suivants :

• **la médiane** est une autre représentation de valeur centrale. Elle se calcule de manière très simple en ordonnant toutes les valeurs de la variable la plus petite, à laquelle est attribué un rang égal à 1, à la plus grande, à laquelle est attribué un rang égal à n. En associant donc un rang de 1 à n à ces valeurs ordonnées, la médiane correspond à la valeur du milieu, c'est-à-dire à la valeur de la variable associée au rang $(n + 1)/2$ si le nombre d'observations est impair, sinon à la moyenne des valeurs associées aux rangs $n/2$ et $n/2 + 1$ si le nombre d'observations est pair. Au final, il y a donc 50 % des valeurs observées qui sont plus petites que la médiane, et 50 % qui sont plus grandes. Elle est moins sensible que la moyenne aux valeurs extrêmes. Lorsque la distribution des valeurs suit une distribution symétrique comme la loi normale, la médiane se confond avec la moyenne. Pour l'exemple de la *figure 2*, la médiane est égale à 49,5 ;

• **l'écart-type** (SD pour *standard deviation*) est la racine carrée de la variance. Pour l'exemple de la *figure 2*, l'écart-type est égal à 10,15 ;

• **les extrêmes** (*range*) correspondent, comme leur nom l'indique, aux valeurs minimales et maximales observées. Pour l'exemple de la *figure 2*, les valeurs varient entre 17 et 78 ans ;

• **l'étendue inter-quartile** (IQR pour *inter-quartile range*) correspond à la différence entre les 25e et 75e percentiles (ou 1er et 3e quartiles) des valeurs observées. Les percentiles s'obtiennent sur la série des observations ordonnées. Par exemple, sur 500 observations ordonnées de la plus petite à la plus grande, les 25e et 75e percentiles correspondent aux 125e et 375e valeurs ordonnées. L'IQR est de plus en plus utilisé dans les représentations graphiques de type Box-Cox ou boîte à moustaches. Pour l'exemple de la *figure 2*, ces 25e et 75e percentiles valent 43 et 56, ce qui donne une valeur d'IQR égale à 13 ;

• **l'intervalle de confiance de la moyenne** (**IC95 %**, *confidence interval*) correspond à un intervalle qui a 95 % de chances de contenir la vraie valeur μ de la moyenne. Il est estimé par la formule suivante : $m \pm 1{,}96\ s / n^{1/2}$. La valeur 1,96 correspond à la valeur x dans la formule de la loi normale centrée réduite $f(x) = 0{,}025$.

À retenir

La loi normale est **représentée** par la moyenne μ et la variance σ^2.
La moyenne μ est **estimée** par m et la variance σ^2 est estimée par s^2.

Variables qualitatives

Il est souvent fait l'assimilation entre des qualités non mesurables et des *variables qualitatives*, mais des variables mesurables peuvent aussi être découpées en classes ou catégories. Dans ce cas, ces variables quantitatives deviennent des variables catégorielles binaires ou ordonnées selon le nombre de classes.

La loi binomiale

Variables à deux classes

Les *variables qualitatives à deux classes* sont appelées dichotomiques ou binaires, par exemple ; homme/femme, ménopause oui/non, succès/échec. Les statistiques qui résument les variables de ce type sont les fréquences et les taux (ou les pourcentages). Par exemple, si on considère la variable X égale au nombre de succès sur un échantillon de taille n, le rapport $p = X / n$ est une estimation du taux de succès. Par définition, la valeur de p varie entre 0 et 1, car X varie entre 0 et n.

Une loi binomiale a deux paramètres n et p. C'est une loi de probabilité qui correspond à l'expérience suivante : on renouvelle n fois de manière indépendante une « épreuve de Bernoulli » de paramètre p [expérience aléatoire à deux issues possibles, généralement dénommées respectivement « succès » et « échec », la probabilité d'un succès étant p, celle d'un échec étant son complément $q = (1 - p)$]. On compte alors le nombre de succès obtenus à l'issue des n épreuves et on appelle X la variable aléatoire correspondant à ce nombre de succès. Il s'agit d'une loi discrète, qui s'applique pour les variables aléatoires discontinues, et qui peut prendre $n + 1$ valeurs entières (0, 1, 2, ... k, ... n) ; par exemple, le nombre de filles dans une famille de 3 enfants. La loi binomiale permet d'estimer la probabilité que le nombre de succès X soit égal à une valeur spécifique. Elle est obtenue par la formule suivante :

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

avec $0 \leq k \leq n$.

La moyenne et la variance sont égales respectivement à : $\mu = n\,p$ et $\sigma^2 = n\,p\,(1 - p)$.

Le taux de succès *(p)* est estimé par la somme de toutes les valeurs observées x_i (de $i = 1$ à n) divisée par le nombre total d'observations :

$$p = \sum_{i=1}^{n} x_i / n$$

où la variable x_i prend la valeur 1 en cas de succès et 0 en cas d'échec. C'est une formulation semblable à celle de la moyenne.

Quand le nombre de répétitions n de l'expérience augmente, on peut approximer la loi binomiale par une loi normale.

À retenir

Ne pas confondre le taux de succès avec le pourcentage de succès, qui vaut 100*p.
Un intervalle de confiance à 95 % correspond à la moyenne (ou probabilité) plus ou moins deux écarts-types.

Variables à plus de deux classes

Dans le cas d'une classification TNM d'une tumeur par exemple, la variable qualitative du stade T est représentée en plusieurs « classes » ou « catégories ». Ces variables sont d'abord exprimées en fréquence (nombre de patients appartenant à chaque classe), puis en pourcentage (par ex., % de T1, % de T4) et peuvent être représentées par un diagramme en barre *(figure 3)*.

Les variables à plusieurs catégories peuvent être ordonnées, comme lorsqu'elles expriment le degré d'envahissement dans le cancer du rectum (< 1 mm, 1-2 mm, 2-3 mm, 3-4 mm, etc.), ou non ordonnées dans le cas de l'énumération des différents sites métastatiques par exemple (os, poumon, foie, etc.). Selon la nature de la variable, les tests statistiques ne sont pas les mêmes (*cf.* chapitre I.4, page 28).

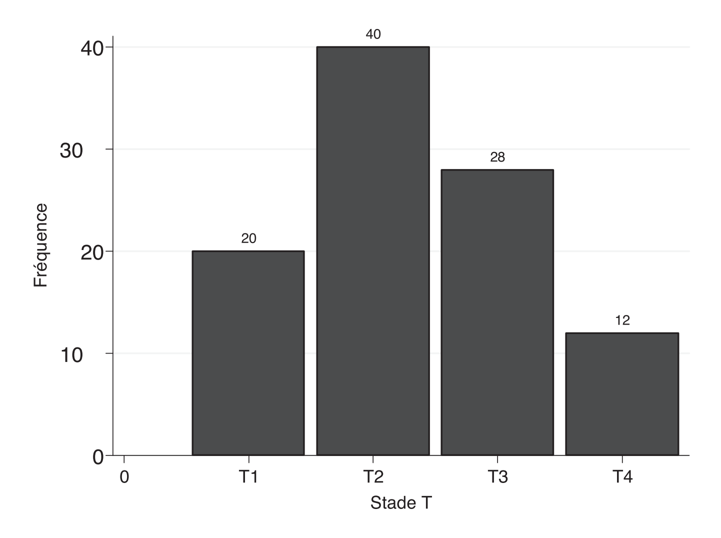


Figure 3. Exemple de diagramme en barres : répartition du stade T (N = 100).

La loi de Poisson

La loi de Poisson a été introduite par Siméon-Denis Poisson (1781-1840) et publiée en 1838 dans son œuvre *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Elle est fondée sur des variables aléatoires qui comptent le nombre d'occurrences discrètes des événements.

Paramètres de la loi de Poisson

La loi de Poisson est souvent utilisée quand la taille de l'échantillon est grande et que le taux d'apparition de l'événement d'intérêt est faible. Elle permet d'estimer la probabilité de survenue du nombre d'événements dans un intervalle donné.

$$P(X = k) = \frac{\lambda^k}{k!} . e^{-\lambda}$$

Par exemple, la tolérance, qui est souvent gradée en catégories ordonnées allant de 0 à 4, peut être décrite par la loi de Poisson. Si le grade moyen de toxicité est de $\lambda = 0{,}50$, en appliquant la formule pour $k = 0, 1, 2, 3$ et 4, on peut s'attendre à observer sur 100 patients, 61 patients avec grade 0, 30 patients avec un grade 1, 8 patients avec un grade 2 et 1 patient avec un grade 3.

Quand λ augmente, la loi de Poisson peut être approximée par la loi normale.

Remarques générales

Notion de mesure

Nous avons vu que les variables quantitatives étaient le résultat d'une mesure. Cela nous permet d'introduire la notion d'erreur de mesure. Chaque mesure est assortie d'un risque d'erreur qui amplifie la variabilité du résultat.

Échantillonnage

En pratique, on ne connaît pas la « vraie » valeur moyenne d'une variable, telle que le poids, dans la population générale, sans la mesurer chez tous les individus. Il faut donc constituer un échantillon représentatif de cette population et mesurer la valeur du poids dans cette « sous-population ». On peut ensuite établir sa distribution et accompagner l'estimation de la moyenne d'un intervalle de confiance, qui fournit des limites inférieures et supérieures à l'intérieur desquelles il y a 95 % de chance que la vraie valeur de la moyenne du poids (dans la population générale) se situe.

Comme il n'est pas faisable d'obtenir des résultats sur toute la population, la constitution d'un échantillon représentatif par tirage au sort (randomisation) est le meilleur moyen d'obtenir des estimations non biaisées (*cf*. chapitre VI.2 « Modalités de randomisation », page 344).

À retenir

Si on prend 100 échantillons représentatifs d'une population et que, pour chaque échantillon, on calcule l'intervalle de confiance de la moyenne, 95 de ces 100 intervalles de confiance contiendront la vraie valeur de la moyenne de cette population.

Pièges

Transformation des variables

Il est important de bien définir le type des variables en séparant donc les variables quantitatives des variables qualitatives, afin d'utiliser le test statistique approprié. Le codage des variables qualitatives à l'aide de valeurs numériques, qui est parfois nécessaire en situation de soins, fait courir le risque de les assimiler à des variables quantitatives et d'appliquer un test inadéquat (par exemple dans l'utilisation d'une échelle de douleur gradant la douleur de 1 à 5 ou de 1 à 10).

Il est aussi possible de transformer des variables quantitatives en variables qualitatives. L'exemple le plus fréquent est l'expression de la variation du poids des patients au cours d'un traitement en fonction de classes liées à l'importance de l'amaigrissement ; perte de plus ou moins de 10 % du poids du corps. Là encore, les variables apparaissent comme qualitatives alors qu'elles sont issues de mesures quantitatives. C'est le cas avec les classes d'âge ou avec un dosage normal ou anormal. Cette transformation est souvent utile mais il faut savoir qu'elle se fait au prix d'une perte de l'information complète (entre la valeur et sa catégorie).

Une autre source de confusion entre les variables quantitatives discrètes et qualitatives ordinales peut apparaître quand il n'y a plus de notion de proportionnalité entre les catégories adjacentes. En effet, la description d'une maladie à l'aide du TNM sous-entend une notion d'aggravation mais qui n'est chiffrable qu'artificiellement. L'écart entre le stade I et le stade II n'est pas mesurable au sens propre du terme par une « unité de mesure ». Ces données sont du domaine du descriptif et sont évaluées par un expérimentateur selon des critères histologiques (envahissement ganglionnaire, taille, etc.). À l'opposé, passer d'un à deux enfants atteints d'une maladie dans une fratrie implique cette notion de proportionnalité.

Conclusion

Avant de commencer une analyse statistique, il faut prendre connaissance des variables puis en déterminer le type (variable qualitative ou quantitative) et la nature (ordonnée, continue ou discrète), afin de choisir la manière juste d'estimer les paramètres et d'appliquer les tests statistiques appropriés. Il convient aussi de respecter les conditions d'application des tests statistiques qui s'en suivent, de vérifier la distribution (normalité par exemple) des données, l'importance des effectifs (petits ou grands) et l'appariement ou non des données. Ces notions seront reprises dans le chapitre I.4 sur le choix du test statistique (*cf.* page 28).

En cancérologie, un certain nombre d'études fait appel à des variables de santé subjectives dans leurs objectifs principaux ou secondaires. En effet, le critère de jugement « idéal » de la meilleure réponse objective ou de la plus longue survie n'est pas toujours le plus pertinent selon les populations considérées. La prise en compte d'informations subjectives telles que la qualité de vie des patients, leur douleur, le bénéfice clinique est nécessaire et fréquemment présente en recherche comme dans le soin. Ces variables peuvent être évaluées par le clinicien en fonction des données de l'interrogatoire – par exemple, pour l'état général : Index de Karnofsky ou score OMS (Organisation mondiale de la santé) – ou estimées par le patient lui-même en fonction d'échelles ou de questionnaires qui lui sont proposés.

Statistiques descriptives

A. Laplanche, E. Benhamou

Ce chapitre sur les statistiques descriptives est divisé en deux parties qui correspondent aux deux volets des résultats d'une publication : tout d'abord la description des sujets inclus et ensuite la présentation des résultats. Certaines notions du chapitre précédent sont reprises pour montrer leur utilisation sur des exemples réels.

Description des sujets inclus

La publication d'une étude de recherche clinique comporte toujours un premier tableau, qui décrit les caractéristiques initiales des sujets inclus et permet au lecteur de connaître la population étudiée et de savoir à qui extrapoler les résultats. Il s'agit le plus souvent d'un tableau à une colonne (étude pronostique, essai de phase II) ou à deux colonnes (essai randomisé en deux groupes, étude cas-témoin). Dans le cas d'un essai randomisé, le premier tableau comporte généralement plusieurs colonnes car il est usuel de présenter les caractéristiques initiales des patients de chaque groupe de traitement, ce qui permet au lecteur de vérifier qu'elles ne diffèrent pas. Présenter une dernière colonne avec les valeurs de *p* associées au test statistique de comparaison de celles-ci n'a pas de sens ici, car le traitement ayant été attribué par tirage au sort, les écarts observés pour les caractéristiques initiales des patients ne sont, par définition, que le fruit du hasard. Néanmoins, il faut rester vigilant sur la répartition de facteurs pronostiques entre les deux groupes.

Exemple 1

Ci-après figure le *tableau I* d'une étude pronostique rétrospective ayant pour but de quantifier et décrire les événements thromboemboliques observés après une chimiothérapie à base de cisplatine. Cent patients de sexe masculin traités à l'Institut Gustave-Roussy entre janvier 1994 et mai 1998 pour une tumeur germinale ont été étudiés [1].

Dans ce tableau, sont présentées des variables qualitatives ou catégorielles (histologie, site du primitif, stade, etc.) et des variables quantitatives ou continues (âge, poids, taille, etc.). En médecine, la plupart des variables catégorielles proviennent de mesures quantitatives. Ainsi, un

Tableau I. Caractéristiques initiales des patients.	
	Patients (n = 100)
Âge (années), médiane (extrêmes)	30 (16-62)
Histologie, n (%)	
tumeur germinale non séminomateuse	86 (86 %)
séminome	14 (14 %)
Site de la tumeur primitive, n (%)	
testicule	86 (86 %)
rétropéritonéal	7 (7 %)
médiastinal	7 (7 %)
Stade de la tumeur primitive, n (%)	
localisé au testicule	18 (18 %)
rétropéritonéal	42 (42 %)
supradiaphragmatique ou métastatique	40 (40 %)
Classification IGCCCG, n (%)	
bon pronostic	65 (66 %)
pronostic intermédiaire	4 (4 %)
mauvais pronostic	29 (30 %)
Fumeur, n (%)	41 (63 %)
Poids (kg), médiane (extrêmes)	73 (47-99)
Taille (cm), médiane (extrêmes)	176 (161-193)
Index de performance (état général), n (%)	
0	76 (76 %)
1	17 (17 %)
2	6 (6 %)
3	1 (1 %)
Chambre implantable ou cathéter, n (%)	43 (43 %)
Neutropénies fébriles, n (%)	19 (19 %)
Chirurgie post-chimiothérapie, n (%)	39 (39 %)
Traitement préventif des thromboses, n (%)	12 (12 %)

marqueur biologique est d'abord quantitatif avant d'être dichotomisé en normal ou anormal. De même, chacune des quatre classes de la réponse tumorale (réponse complète, réponse partielle, stable, progression) est définie à partir de la différence de mesures quantitatives.

• Les **variables qualitatives** sont décrites par leur effectif et pourcentage. Il est recommandé de ne pas fournir des informations redondantes, comme les effectifs et pourcentages de femmes si

l'on a présenté ceux des hommes ou, ici, ceux des non-fumeurs qui se déduisent immédiatement de ceux des fumeurs. En revanche, nous conseillons d'indiquer l'effectif et le pourcentage de chaque variable car cela permet de repérer les données manquantes. Ainsi pour la caractéristique « fumeur », les informations : 41 fumeurs (63 %) permettent de savoir que cette donnée manquait pour 35 patients. En effet, ces résultats ont été observés sur 65 sujets (41/0,63 = 65) et non sur 100. Une autre possibilité est d'indiquer dans le tableau dans une ligne supplémentaire le nombre de données manquantes.

À retenir

Veiller à ne pas reporter les nombreuses décimales fournies systématiquement par les logiciels car elles nuisent à la lisibilité des résultats.

• Les **variables qualitatives ordonnées** (index de performance, par exemple) ne doivent pas être traitées comme des variables quantitatives. Il est recommandé de présenter les effectifs et pourcentages de chaque catégorie.

• Les **variables quantitatives** doivent être présentées sans oublier de préciser leur unité : âge en année, taille en cm, etc. Elles sont décrites par deux indicateurs : un indicateur de valeur centrale (moyenne ou médiane) et un indicateur de dispersion (écart-type ou extrêmes).

Dans le *tableau I* sont présentés la médiane et les extrêmes de chaque variable quantitative. Un âge médian de 30 ans signifie que 50 % des sujets de l'étude ont moins de 30 ans et 50 % plus de 30 ans. Les extrêmes *(range)* sont l'âge minimum (16 ans) et l'âge maximum (62 ans). On aurait tout aussi bien pu indiquer l'âge moyen (32 ans) et son écart-type (9 ans).

Lorsque la variable a une distribution normale – c'est la distribution en cloche symétrique autour de la moyenne (*cf*. chapitre précédent) –, **médiane** et **moyenne** sont proches et présenter l'une ou l'autre est indifférent. Il faut cependant savoir que la médiane n'est pas influencée par des valeurs extrêmes ou erronées. Ainsi si l'on a codé 620 au lieu de 62 pour la valeur de l'âge le plus élevé, la moyenne devient 37 ans alors que la médiane reste inchangée. On choisira la médiane pour les variables dont la distribution n'est pas normale (titres d'anticorps, variable monétaire, par exemple) et en cas de petits échantillons (classiquement moins de 30 sujets).

Une mesure de dispersion bien connue est l'**écart-type** qui, derrière une formule un peu barbare, cache un concept très simple. Si l'on reprend l'exemple de l'âge, il représente la dispersion des âges des 100 sujets de l'échantillon autour de leur moyenne (32 ans). On calcule donc l'écart d'âge de chaque sujet par rapport à la moyenne, comme indiqué en colonne 2 du *tableau II*.

La dispersion autour de la moyenne est d'autant plus grande que les écarts individuels sont grands. Comme on ne peut pas sommer directement ces écarts – cette somme serait nulle puisque environ la moitié de ceux-ci sont positifs et la moitié sont négatifs –, c'est leur carré (colonne 3) dont on fait la somme. Enfin, cette somme étant d'autant plus grande qu'il y a beaucoup d'observations, on en fait la moyenne en la divisant par le nombre de sujets moins un, ici 99. La quantité obtenue (82,21) est la **variance**. Cependant, on présente rarement la variance dans le premier

Tableau II. Exemple de calcul de la variance de l'âge.			
	①	②	③
	Âge	**Âge-moyenne**	**(Âge-moyenne)^2**
	16	**- 16**	**256**
	16	**- 16**	**256**
	17	**- 15**	**225**
	...	**...**	**...**
	31	**- 1**	**1**
	32	**0**	**0**
	33	**+1**	**1**
	34	**+2**	**4**
	...	**...**	**...**
	50	**+ 18**	**324**
	62	**+ 30**	**900**
Somme	*3 200*	*≈ 0*	*8 139*
Moyenne	*32*		
Variance			*82,21*
Écart-type			*9,07*

tableau de l'article. En effet, elle n'est pas directement interprétable car son unité est celle de la variable au carré (année^2). C'est la raison pour laquelle on présente l'écart-type (SD pour *Standard Deviation*), qui est la racine carrée de la variance et a donc la même unité que la variable. Ici SD = 9 ans. Il ne faut pas confondre SD et SEM *(Standard Error of the Mean)* : SD est l'écart-type des observations alors que SEM est l'écart-type de la moyenne. SEM s'obtient en divisant SD par $n^{1/2}$, n étant la taille de l'échantillon. L'une ou l'autre des deux informations peut être utile selon l'usage que l'on veut en faire.

En conclusion, on peut présenter soit la médiane et les extrêmes, soit la moyenne et l'écart-type (SD), mais il est indispensable de préciser quelles sont les mesures présentées dans le tableau !

Une propriété d'une variable distribuée selon la loi normale est que 95 % des observations se situent dans l'intervalle compris entre la moyenne moins 1,96 écart-type et la moyenne plus 1,96 écart-type.

À retenir

Le chiffre 1,96 est très souvent approximé par 2 (c'est ce qui a été fait plus loin avec moyenne ± 2 SD ou moyenne ± 2 SEM).

Dans notre étude, le poids moyen (SD) des sujets est de 73 kg (12), ce qui permet de définir l'intervalle : [73 - 2*12 ; 73 + 2*12] = [49 ; 97].

On s'attend donc à avoir 95 % des poids dans cet intervalle, donc 5 % en dehors, soit 2,5 % au-delà de 97 kg et 2,5 % inférieurs à 49 kg. C'est ce que l'on retrouve sur les données de notre exemple : 3 patients pesaient plus de 100 kg, 1 patient pesait 47 kg et 1 patient pesait 48 kg. Par analogie, on peut calculer l'intervalle qui contient 50 %, 75 %, 99 %... des observations en remplaçant 1,96 par le chiffre correspondant, que l'on peut trouver dans les tables de la loi normale.

Exemple 2

Dans la vie quotidienne, on se sert souvent sans le savoir de cette propriété de la **loi normale**, par exemple en utilisant la courbe de poids qui figure dans le carnet de santé des enfants. La courbe ci-dessous *(figure 1)* permet de savoir qu'une petite fille de 1 an qui pèse 12 kg fait partie des 3 % d'enfants classés en « surpoids ».

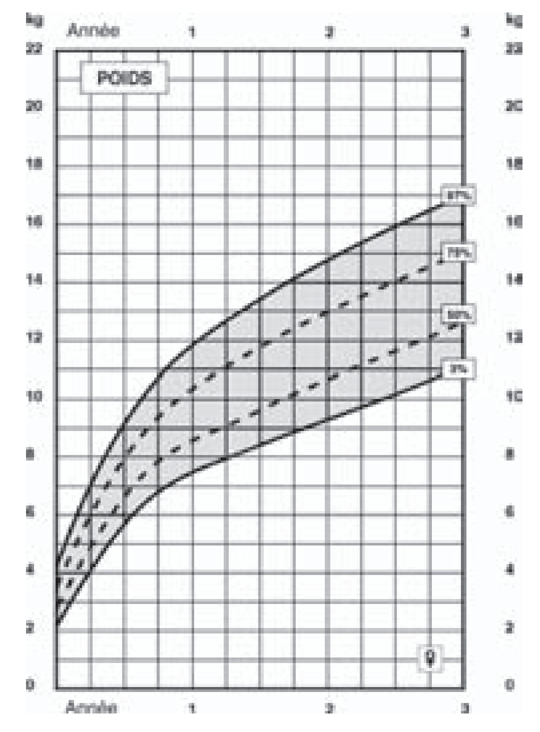


Figure 1. Courbe de poids des filles de 0 à 3 ans.

Présentation des résultats

La présentation des résultats dépend du type de critère étudié. Nous introduirons la notion d'intervalle de confiance et discuterons la présentation de critères quantitatifs. Le lecteur se reportera au chapitre III.1 « Données de survie » (*cf.* page 129) pour la présentation des résultats de survie.

Intervalle de confiance

Revenons à l'exemple 1 ayant pour but de quantifier et décrire les événements thrombo-emboliques (TE) observés après une chimiothérapie à base de cisplatine. Sur l'échantillon de 100 sujets décrit au *tableau I*, nous avons observé 19 TE. Ce pourcentage observé est soumis aux fluctuations d'échantillonnage : avec un autre échantillon de 100 sujets, nous aurions très bien pu observer 17 %, 21 %, voire 30 % de TE ! Un calcul des probabilités permet « d'encadrer » le pourcentage observé 19 % dans un intervalle dans lequel on a « suffisamment confiance » que se situe « le pourcentage vrai inconnu ». On a « suffisamment confiance » si cet intervalle a 95 chances sur 100 de contenir le vrai pourcentage, d'où son nom : « **intervalle de confiance à 95 %** » (IC95 %). Ici, on a IC95 % : [12 % ; 28 %].

Au total, avec un risque de 5 %, sur 100 estimations de pourcentages, on donnera en moyenne 5 « fourchettes » fausses, c'est-à-dire 5 intervalles ne contenant pas la vraie valeur. Si l'on veut réduire le risque de se tromper, par exemple à 1 %, on donnera l'IC à 99 % (ici [10 % ; 31 %]) qui est plus large, donc moins informatif. Et si l'on ne veut jamais se tromper, on donnera un intervalle qui va de 0 à 1 (et ne présente plus aucun intérêt).

On calcule l'IC95 % à l'aide des formules suivantes sur un échantillon de *n* sujets pour un pourcentage *(p)* et une moyenne *(m)* :

$$p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

$$m \pm 1.96\frac{SD}{\sqrt{n}} = m \pm 1.96 \times SEM$$

> **Remarque**
>
> Il existe une formule exacte à utiliser en cas de petits échantillons ou pour des pourcentages proches de zéro ou de un.

En règle générale, pour améliorer la précision d'une estimation, c'est-à-dire diminuer la largeur de son intervalle de confiance, on augmente le nombre de sujets de l'échantillon. Ainsi, dans notre exemple :

$n = 100$	$p = 19$ %	IC95 % [12 % ; 28 %]
$n = 1\,000$	$p = 19$ %	IC95 % [17 % ; 22 %]

> **Remarque**
>
> Calculer un intervalle de confiance suppose de réaliser une « inférence » à la population dont est issu l'échantillon observé. Cela n'a donc de sens que pour le critère étudié, et non pour la description des caractéristiques des sujets de l'échantillon.

Exemple 3

Dans une étude épidémiologique cas-témoin sur le cancer de la vessie, les adduits[1] à l'ADN, **critère quantitatif**, ont été mesurés dans le tissu vésical sain de 59 cas et de 45 témoins [2].

Les médianes (extrêmes) des adduits sont égales à 18 (8-63) et 15 (9-46) chez les cas et les témoins respectivement. Les moyennes (SD) sont égales à 19 (9) et 18 (7) chez les cas et les témoins respectivement. Pour représenter graphiquement les niveaux d'adduits des cas et des témoins, on peut choisir l'une des trois figures ci-dessous : *figure 2a* présentant les boîtes à moustaches ou « Box plots », *figure 2b* présentant la moyenne ± 2 SD et *figure 2c* présentant la moyenne ± 2 SEM (**attention** l'échelle des trois figures n'est pas la même).

La représentation en Box plot est celle qui contient le plus d'information. Dans tous les cas, il est recommandé d'avoir une légende très explicite afin que le lecteur puisse savoir quelle est l'information représentée (les barres verticales autour de la moyenne peuvent correspondre à 1 ou 2 SD ou à 1 ou 2 SEM…).

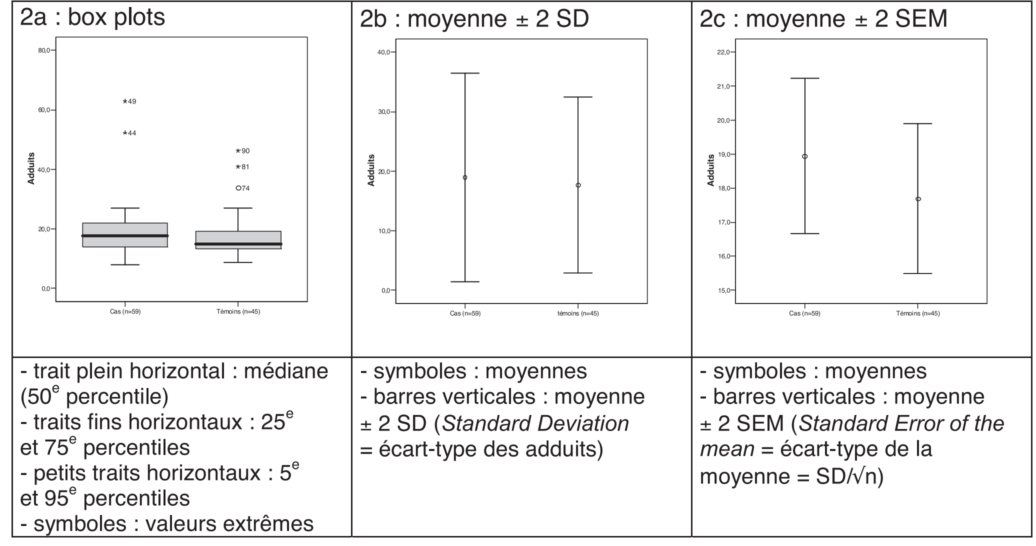


Figure 2. Adduits à l'ADN.

1. Les adduits représentent les molécules de l'agent carcinogène liées de façon covalente d'une base azotée de l'ADN. S'ils ne sont pas éliminés par des processus de réparation de l'ADN, les adduits sont susceptibles de provoquer des mutations.

Recommandations CONSORT

Pour insister sur l'importance des notions présentées dans ce chapitre, nous faisons le lien avec des recommandations internationales : les recommandations CONSORT *(Consolidated Standards of Reporting Trials)* [3]. Il s'agit de 25 recommandations (25 items) issues d'un groupe d'éditeurs internationaux qui ont pour but d'améliorer la qualité des publications d'essais cliniques. Le 15e item concerne la présentation des « caractéristiques initiales démographiques et cliniques de chaque groupe ». Le 17e item concerne la présentation du « résumé des résultats de chaque groupe ». Un exemple de présentation de chaque item, exemple issu de la littérature internationale récente, est donné dans l'article *princeps* de CONSORT [3].

Références

1. Piketty AC, Flechon A, Laplanche A, *et al*. The risk of thrombo-embolic events is increased in patients with germ-cell tumours and can be predicted by serum lactate dehydrogenase and body surface area. *Br J Cancer* 2005 ; 93 : 909-14.

2. Benhamou S, Laplanche A, Guillonneau B, *et al*. DNA adducts in normal bladder tissue and bladder cancer risk. *Mutagenesis* 2003 ; 18 (5) : 445-8.

3. Moher D, Hopewell S, Schulz KF, *et al*. CONSORT 2010 Explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *BMJ* 2010 ; 340 : c869.

Compréhension des tests statistiques

B. Pereira, S. Thezenas

Ce chapitre sur la compréhension des tests d'hypothèse détaille la démarche expérimentale dans l'évaluation des stratégies thérapeutiques : seront abordés la notion des hypothèses nulle et alternative, les risques d'erreur alpha et bêta, ainsi qu'une section sur le calcul de la taille de l'échantillon (nombre de sujets nécessaires).

Les principes de l'expérimentation

À partir de la mise en place d'un protocole d'étude où la décision est prise d'évaluer tel ou tel traitement, et sur quels patients, il reste la question essentielle à résoudre : comment comparer des groupes ? Aussi, le problème qui nous intéresse essentiellement dans les essais thérapeutiques va au-delà de la description d'un ou deux groupes de patients, et il s'agit surtout de comparer les résultats des traitements entre groupes. Par exemple, si un médicament A donne 24 guérisons sur 38 patients, et un médicament B 28 guérisons sur 35 patients, peut-on conclure que le médicament B est meilleur que le médicament A ? En nombre absolu de guérisons, B est légèrement supérieur à A ! On sait que même si l'on compare deux groupes traités par deux médicaments strictement équivalents (voire même deux groupes traités exactement par le même médicament), on n'obtiendra que rarement la même proportion de guérisons d'un groupe à l'autre. Les véritables questions qui se posent sont en fait : 1) la différence observée est-elle simplement due au hasard seul (on parle alors de fluctuation d'échantillonnage) ? ou 2) est-il possible avec une certaine assurance d'affirmer la supériorité d'un traitement sur l'autre [1] ?

Quel que soit le test statistique considéré, il sera toujours appliqué afin de mesurer les effets observés et ainsi estimer la probabilité que les résultats peuvent ou non être obtenus simplement par le fait du hasard [2, 3].

Tous les tests statistiques obéissent à une même règle résumée comme suit :
- définir les hypothèses H_0 et H_1 ;
- fixer un seuil α du risque de première espèce (défini ci-dessous) ;
- choisir le test approprié (*cf.* chapitre I.4 « Choix du bon test statistique », page 28) ;
- déterminer une région de rejet, c'est-à-dire l'ensemble des valeurs de la statistique de test pour lesquelles l'hypothèse nulle H_0 est rejetée (acceptation par défaut de H_1) ;

- calculer la valeur du test statistique à partir d'un échantillon aléatoire ;
- prendre la décision et inférer la véracité de la théorie.

Hypothèses nulle et alternative

Dans le cadre d'une comparaison pour évaluer la supériorité d'un traitement par rapport à un autre, le test de signification commencera par une supposition qu'il n'y a pas d'effet de traitement : l'hypothèse nulle H_0 (pas de différence). Désireux d'examiner si cette hypothèse H_0 est plausible, on construit un test de signification sur les données observées pour mesurer s'il y a suffisamment de preuves contre cette hypothèse. Il s'agit du principe de la réfutation proposé par Karl Popper afin d'étudier la causalité et dont l'objectif est de prouver que notre hypothèse H_0 est fausse. Ce principe est ici appliqué pour les tests d'inférence statistique.

Quelle est la signification du terme « hypothèse nulle » ?

- On ne peut pas démontrer la vérité de la théorie à partir de la vérité du fait. On cherchera donc à démontrer qu'elle est fausse (négation d'une hypothèse de recherche = réfutation).

- Hypothèse sur la base de laquelle est définie la distribution de probabilité (théorie de la décision).

- Autant la spécification de l'hypothèse nulle est précise (différence égale à zéro), autant l'hypothèse alternative est imprécise (différence non égale à zéro).

Construire le test de H_0 contre H_1 (dite hypothèse alternative) revient donc à adopter une règle de décision qui amène à rejeter ou ne pas rejeter H_0 [4].

Tous les tests statistiques répondent à une procédure composée de deux éléments :
- un tableau de prise de décision qui inclue des risques d'erreur *(tableau I)* ;
- une série d'étapes pour déterminer les tests statistiques qu'il faut utiliser selon la situation (comparaison de moyennes, de proportions, de taux de survie, l'adéquation à une loi de probabilité, etc.).

Risques d'erreur

Au terme de toute étude, la décision de rejeter ou non l'hypothèse nulle H_0 est fondée sur les données observées et leur comptabilité ou non par rapport à l'hypothèse nulle [5] :

Compte tenu d'une hypothèse nulle qui se pose en affirmant l'absence de différence, quatre situations sont envisageables selon que H_0 soit vraie ou fausse. Deux des situations *(tableau I)* correspondent à des risques d'erreur engagés sur toute prise de décision reposant sur un test statistique. Ces deux erreurs sont :
- risque α (erreur de type I ou de première espèce) : on prétend que le nouveau traitement B est meilleur que le traitement de référence A, alors qu'en réalité c'est faux ! (on utilise également le

terme de « faux positifs »). Si on croit ces résultats et qu'on change les pratiques, on prend donc le risque de traiter des patients avec B alors qu'il n'est pas plus efficace que A. Par conséquent, plus on veut diminuer ce risque, plus α doit être petit ;

- risque β (erreur de type II ou de seconde espèce) : on affirme que le nouveau traitement B n'est pas meilleur que le traitement de référence A, alors que c'est faux ! Si on ne change pas les pratiques, on prend le risque de ne pas traiter des patients avec un traitement plus efficace (on utilise également le terme de « faux négatifs »).

Les graphiques de la *figure 1* caractérisent le choix parfois difficile entre deux décisions. Deux distributions de la loi normale sont présentées ; celle de gauche représentant la distribution centrée-réduite (de moyenne zéro et variance 1), et celle de droite représentant la distribution réduite (de moyenne 3 et variance 1) (*cf.* chapitre I.1 « Distributions statistiques », page 3). Ces figures représentent les distributions de la statistique sous les hypothèses nulle et alternative. Lors de la réalisation d'un test statistique, il faut choisir entre ces deux hypothèses : H_0 : « les moyennes μ_A et μ_B sont égales » ($H_0 : \mu_A - \mu_B = 0$) contre H_1 : « il y a une différence entre les moyennes μ_A et μ_B » ($H_1 : \mu_A - \mu_B = 3$). Le résultat du test statistique se résume par une seule valeur qu'on va situer sur l'axe des ordonnées pour voir où il se place par rapport à l'une ou l'autre hypothèse. Par exemple, selon des règles établies *a priori* à partir d'une valeur de Z (valeur centrée réduite de la loi normale (0,1)) supérieure à 1,645, on prendra la décision de rejeter l'hypothèse nulle (pas de différence), car on voit que l'on s'éloigne de cette hypothèse en faveur de l'hypothèse alternative et on pourra conclure que les moyennes sont statistiquement différentes.

Tableau I. Risques d'erreurs associés aux règles de décisions.

	Décision	
Réalité	***Non-rejet de H_0***	***Rejet de H_0***
H_0 vraie	Décision juste : 1-α	Erreur de 1re espèce : α
H_1 vraie	Erreur de 2e espèce : β	Décision juste : 1-β

Pour minimiser le risque d'une mauvaise décision, ces deux risques (α, β) doivent être les plus faibles possibles. En pratique, on en fixe le niveau de manière subjective en convenant qu'il est globalement plus grave d'affirmer une différence ou une relation qui n'existe pas, que d'ignorer une relation ou une différence existante (généralement 1 % ou 5 % pour α et 5 %, 10 % ou 20 % pour β). Il s'agit des risques maximums que l'on est prêt à prendre. Par exemple, pour une valeur de p calculé de 9 %, il y a donc 9 % de risque de dire que l'hypothèse H_0 est fausse alors qu'en réalité elle est vraie. Si le risque maximum acceptable α a été fixé *a priori* à hauteur de 5 %, on décidera alors de ne pas rejeter l'hypothèse nulle. En revanche, si on avait choisi un niveau de significativité de 10 %, alors on conclura à une différence statistiquement significative. Plus le niveau fixé *a priori* est élevé, plus on a de chances de faire apparaître une relation significative, mais plus nombreux seront les résultats faussement positifs.

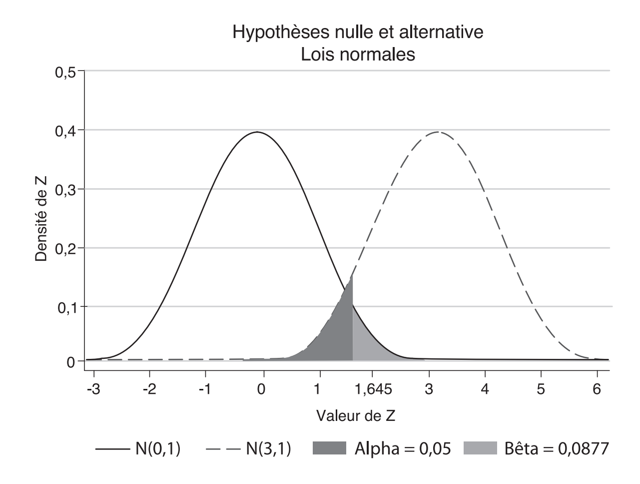


Figure 1. Représentation des risques d'erreur de 1re et 2e espèces.

À retenir

Fixer un seuil, par exemple de 5 %, pour le risque de première espèce α permet d'assurer que si l'expérience est répétée plusieurs fois, l'hypothèse nulle sera rejetée à tort 5 fois sur 100. L'utilisation de ce seuil permet de rejeter l'hypothèse nulle à tort dans 5 % des cas, ce qui correspond aux cas de faux positifs.

Lorsque l'hypothèse nulle n'est pas rejetée, cela ne signifie pas qu'elle soit vraie, mais seulement que le test utilisé n'a pas pu la rejeter au vu des résultats de l'expérience. Cela est très important dans l'interprétation souvent erronée de certains articles devant une absence de différence entre deux traitements, qui prétendent à tort l'équivalence des deux traitements.

Puissance statistique

La notion de puissance d'un test statistique est définie à partir du risque de seconde espèce β et la puissance est égale à 1-β. Par analogie avec ce qui a été abordé précédemment, la puissance correspond à la probabilité que l'on admette à raison l'existence d'un résultat significatif. Autrement dit, la puissance d'un test statistique peut être vue comme une mesure de la probabilité que le test produise un résultat statistiquement significatif pour une vraie différence donnée [6].

Considérons un essai dans lequel seulement quelques patients sont recrutés. Même s'il y a une vraie différence thérapeutique (c'est-à-dire hypothèse H_0 fausse), il est peu probable que l'on obtienne suffisamment de preuves contre l'hypothèse nulle ; les résultats ne seront pas considérés comme convaincants. On dit alors que la puissance est faible.

Ainsi, dans un essai avec très peu de patients, on ne pourra détecter que les grandes différences (flagrantes) et manquer de détecter les différences médicalement importantes, même quand elles sont présentes.

A contrario, dans un essai avec beaucoup de patients, on pourra détecter des petites différences (valeur de *p* significative). Par exemple, pour détecter une différence significative entre des taux de succès de 30 % et 50 % par le test du chi-2 (*cf.* chapitre I.4, page 28), 30 patients par groupe n'est pas suffisant (p = 0,114) alors qu'avec 100 patients la différence devient significative *(tableau II)*.

À retenir

La puissance est liée à la force de la relation ou de la différence testée, à la taille de l'échantillon et au risque de première espèce α.

Taille de l'échantillon

Il est largement reconnu que le nombre de sujets dans de nombreux essais thérapeutiques randomisés est vraiment trop faible. Ainsi, parmi 110 essais « négatifs » publiés, 71 essais ont affirmé des phrases du style *« No significant reduction in mortality »*, mais la moitié de ces essais avait une puissance inférieure à 60 % de détecter un bénéfice thérapeutique aussi grand que 50 % [7]. Ainsi donc, beaucoup de traitements sont rejetés car jugés inefficaces à partir d'essais négatifs, potentiellement par manque de puissance alors que les effets cliniques peuvent être importants (bénéfice jusqu'à 50 % !).

Pour obtenir une estimation du nombre de sujets nécessaires dans un essai où les patients sont randomisés entre deux options thérapeutiques, il faut d'abord identifier le critère de jugement qui sera considéré comme le critère principal pour comparer les deux groupes [8].

Tableau II. Valeurs des tests statistiques ayant des taux observés de 30 % et 50 %.

Nombre de patients	Succès du groupe 1	Succès du groupe 2	*p*
60 (30 par groupe)	9 (30 %)	15 (50 %)	0,114
100 (50 par groupe)	15 (30 %)	25 (50 %)	0,041

Source : d'après [9].

• Si l'on s'intéresse à la comparaison **de taux de réponses**, le critère est considéré binaire et la taille de l'échantillon sera calculée dans le contexte de la comparaison de deux pourcentages (un autre exemple de pourcentage : le taux de complications à 30 jours).

• Si l'on s'intéresse à la comparaison d'**un changement de la tension artérielle**, le critère est considéré continu et la taille de l'échantillon sera calculée dans le contexte de la comparaison de deux moyennes.

• Si l'on s'intéresse à la comparaison **des temps jusqu'à un événement critique**, comme le délai jusqu'à la progression, la rechute ou le décès, le critère est un taux d'événement au cours du temps et la taille de l'échantillon sera calculée dans le contexte de la comparaison des courbes de survie ou des fonctions de risque.

Pour déterminer le nombre de sujets N nécessaire pour un essai clinique, il faut connaître les paramètres suivants :

- l'effet thérapeutique que l'on veut être capable de détecter (Δ) ;
- la variabilité dans les données (l'écart-type du critère de jugement, *cf.* chapitre I.1, page 3) ;
- α, la valeur de p souhaitée ;
- 1-β, la puissance du test.

À partir de ces paramètres, on peut calculer le nombre de patients nécessaires pour détecter une différence Δ avec une probabilité suffisamment élevée (puissance d'au moins 80 %) si une vraie différence existe.

Quelquefois l'estimation de la taille de l'échantillon peut être calculée avec des données précises accompagnées par des informations détaillées sur les taux de base et la variabilité. Mais, on peut être dans la situation de manque de données sur le sujet, ce qui entraîne des difficultés dans le calcul de la taille de l'échantillon. Dans ce cas, une étude pilote peut permettre d'obtenir des résultats qui permettront d'obtenir les paramètres nécessaires pour estimer la taille d'une étude plus large. En fait, l'estimation de la taille de l'échantillon et la considération de la puissance ont une importance fondamentale dans le stade de planification de l'étude [10].

Interprétation des résultats

Les résultats issus d'un test statistique de signification sont exprimés par une valeur de p (*p-value*), communément appelé « petit p » (*cf.* chapitre I.5 « Quand dit-on qu'une différence est statistiquement significative ? », page 47). Les preuves contre l'hypothèse nulle sont mesurées par cette valeur de p qui correspond à la probabilité d'obtenir les données telles qu'on les a observées ou des données plus extrêmes si l'hypothèse nulle était vraie (probabilité de rejeter H_0 à tort). Lors du résultat du test statistique de l'hypothèse nulle au niveau α, cette hypothèse est rejetée si la probabilité p est inférieure au seuil α fixé au préalable [11].

Si la valeur de p est suffisamment petite, il est peu vraisemblable que cette différence soit due au hasard, alors on rejette l'hypothèse nulle (la différence observée est dite significative). De même, si le test statistique conduit à ne pas rejeter l'hypothèse H_0, la différence n'est pas statistiquement significative. L'usage est de considérer comme statistiquement significatifs les résultats dont le risque qu'ils soient obtenus par hasard dans le cas d'une égalité des traitements est inférieur à 5 % (*p-value* < 5 %). Le choix du risque alpha sera guidé par les options stratégiques disponibles à la fin de l'essai, et parfois le risque d'un faux positif de 5 % peut être considéré comme inacceptable.

On aimerait avoir une certitude dans les décisions prises à la suite d'une étude, or les tests statistiques ne donneront jamais une certitude absolue mais seulement un risque, si nécessaire aussi petit que voulu.

À retenir

Si le test statistique de signification donne une valeur de p non significative, on ne peut pas dire que l'hypothèse nulle est vraie, car on ne peut jamais prouver l'hypothèse nulle, seulement ne pas la rejeter, dans quel cas elle doit être acceptée comme une possibilité.
Si le résultat de l'essai est statistiquement non significatif, alors on ne peut pas dire qu'il n'y a pas de différence entre les traitements. On peut simplement dire qu'il n'y a pas assez de preuves pour rejeter l'hypothèse nulle. L'effet du traitement est tout simplement très faible. La taille de l'échantillon était peut-être trop petite (ou on n'a pas eu de chance).
Si on obtient une valeur de p significative, on ne peut pas dire que l'hypothèse nulle est fausse, car on ne peut jamais prouver l'hypothèse nulle. On le rejette simplement, ce qui ne veut pas dire qu'on ne la croit pas uniquement parce que les résultats sont en défaveur de cette hypothèse.

Si le résultat de l'essai donne une différence statistiquement significative, on ne peut pas dire que l'effet est cliniquement ou médicalement utile. La signification statistique n'indique pas forcément une différence médicale intéressante et importante (on peut alors parler de signification clinique).

Conclusions

Un test statistique repose sur des hypothèses visant à démontrer qu'une relation n'est pas due au hasard. En termes de résultats, les tests statistiques ne donneront jamais une certitude absolue mais seulement un risque, si nécessaire aussi petit que possible. Ces tests d'inférence statistiques n'évaluent que la probabilité d'une variation aléatoire. Il faut donc que le plan d'expérience de l'essai clinique élimine au mieux la variabilité non aléatoire (les biais), pour que les groupes à comparer soient le plus comparables possibles, à part le traitement.

Références

1. Schwartz D, Flamant R, Lellouch J. *L'Essai thérapeutique chez l'homme*. Paris : Flammarion, 1994, 297 pages.

2. Dagnelie P. *Statistique théorique et appliquée. Tome 2 : Inférence statistique à une et à deux dimensions*. Paris et Bruxelles : De Boeck et Larcier, 2006, 736 pages.

3. Dagnelie P. *Statistique théorique et appliquée. Tome 1 : Statistique descriptive et base de l'inférence statistique*. Paris et Bruxelles : De Boeck et Larcier, 2007, 511 pages.

4. Droesbecke JJ. *Éléments de statistique*. Paris : Ellipses, 2001, 576 pages.

5. Huguier M, Flahault A. *Biostatistiques au quotidien*. Paris : Elsevier, 2000, 204 pages.

6. Falissard B, Lellouch J. *Comprendre et utiliser les statistiques dans les sciences de la vie*. Paris : Masson, 2005, 372 pages.

7. Freiman JA, Thomas AB, Chalmers C, Harry Smith Jr M, Roy R., Kuebler RR. The importance of beta, the type II error and samples size in the design and interpretation of the randomized control trial – Survey of 71 negative trials. *N Engl J Med* 1978 ; 299 : 690-4.

8. Schwartz D. *Méthodes statistiques à l'usage des médecins et des biologistes*. Paris : Flammarion, 1996, 314 pages.

9. Kramar A, Paoletti X. Analyses intermédiaires. *Bull Cancer* 2007 ; 94 (11) : 965-74.

10. Fayers P, Machin D. How many patients are necessary? *Brit Journal Cancer* 1995 ; 72 : 1-9.

11. Saporta G. *Probabilité, Analyse des données et statistique*. Paris : Technip, 1990, 622 pages.

Choix du bon test statistique

A.L. Septans, F. Kwiatkowski

Ce chapitre sur le choix du bon test statistique est consacré à la démarche nécessaire pour bien choisir le test statistique selon le plan expérimental, le type de la variable et les hypothèses.

Test du chi-2, test t de Student, test F de Fisher, test de Kolmogorov-Smirnov, test de Kruskal-Wallis, test de Wilcoxon, test de Mann-Whitney, ANOVA, test du log-rank… Comment être sûr de choisir correctement le test qui permettra de répondre à la problématique posée ? Serait-ce chercher une aiguille dans une botte de foin ? Évidemment non, les tests statistiques reposent sur des conditions d'application et des hypothèses de départ relativement simples. La première étape consiste à identifier les différents types et natures des variables que l'on veut étudier (*cf.* chapitre I.1 « Distributions statistiques », page 3). Dans ce chapitre, deux situations sont examinées : la comparaison des résultats d'une variable entre deux ou plusieurs groupes de patients, et l'étude de la liaison (corrélation) entre deux variables.

Différents types d'analyse

Avant de choisir le test statistique, il faut savoir s'il s'agit de comparer plusieurs groupes ou bien d'étudier la liaison entre plusieurs variables. Ensuite, il faut préciser s'il s'agit de séries indépendantes ou appariées. Enfin, il faut choisir entre deux grandes catégories de test : les tests paramétriques et les tests non paramétriques.

Comparaison ou liaisons ?

S'il s'agit d'une **comparaison**, il est légitime de se poser des questions comme : la répartition d'un dosage biologique est-elle la même entre les patients malades et les patients non malades ? Le pourcentage de sujets malades chez les plus de 60 ans est-il le même que chez des sujets plus jeunes ? La variabilité de la concentration d'uracile plasmatique chez des patients homozygotes mutés est-elle comparable à celle des patients hétérozygotes ?

S'il s'agit d'une **liaison**, on rencontrera des questions comme : la clairance est-elle liée à la dose de produit injecté, la toxicité X est-elle liée à la toxicité Y, le sex-ratio des malades est-il lié à leur appartenance à une classe d'âge…

Séries indépendantes ou appariées ?

Deux grandes situations de test doivent être distinguées : celle où la variable à analyser a été obtenue sur deux ou plusieurs groupes de sujets indépendants, par exemple sur deux groupes de patients affectés à un bras de traitement par randomisation, un groupe de patients traités par une nouvelle molécule *versus* un autre groupe traité par placebo…

Dans la seconde situation, la même variable est recueillie au cours du temps sur le même groupe de sujets, et l'on parle alors de mesures appariées. Par exemple, on désire comparer la concentration d'une variable avant et après l'administration d'un médicament : dans un tel cas, chaque sujet est considéré comme son propre témoin.

La différence essentielle du point de vue statistique entre ces deux situations est le fait qu'une corrélation substantielle existe dans le cas des séries appariées – l'effet de la variabilité inter-individuelle n'interfère quasiment plus –, qui ne peut être négligée pour le choix du bon test statistique.

Si les patients appariés sont différents, c'est-à-dire qu'ils sont issus de deux cohortes distinctes mais ont été appariés *a posteriori* sur des critères cliniques par exemple (même âge, sexe, origine, pathologie, etc.), l'appariement n'offre pas les mêmes qualités : des biais de sélection peuvent alors entacher les résultats. Ce cas de figure comprend les études faites sur des sujets traités « avant » une certaine date et d'autres « après » cette même date.

D'un point de vue pratique, ces deux situations concernent d'une part le problème de comparaison de groupes différents – comparer l'efficacité entre deux groupes de patients traités soit par A soit par B –, d'autre part celui de la comparaison d'un même groupe dans deux situations différentes : par exemple, évaluer l'évolution d'un marqueur biologique avant et après le début du traitement.

Tests paramétriques et non paramétriques

Un certain nombre de tests s'appuie sur l'hypothèse que la variable étudiée suit une loi de distribution connue (par ex., loi normale). C'est la connaissance de cette loi qui permet de calculer la statistique de test et ensuite la probabilité de risque associée. Ce type de test est appelé test paramétrique. Un test est dit non paramétrique, lorsque la distribution de la variable étudiée n'est pas conforme à une loi connue ou qu'il n'y a pas suffisamment d'éléments qui permettent de faire cette hypothèse. Les valeurs observées de la variable sont alors remplacées par leur rang de classement après tri. Ces tests sont peu sensibles aux valeurs aberrantes. Ainsi, la plus grande observation aura la même contribution à la statistique de test si elle est à 2 unités ou à 200 unités de distance de l'avant-dernière observation.

À retenir

Plus un échantillon est petit, plus il est sensible aux valeurs aberrantes.
On privilégiera les tests non paramétriques lorsque les échantillons sont petits.

Les tests non paramétriques sont nombreux et peuvent être adaptés à plusieurs situations. Ci-dessous *(figure 1)*, l'illustration de l'effet d'une valeur aberrante sur le lien existant entre deux variables : le poids et l'âge.

Dans cet exemple, on a volontairement choisi une population disparate : 9 personnes ont un âge compris entre 15 et 23 ans tandis qu'une personne a 70 ans (point isolé à droite). Si l'on étudie le lien âge/poids avec un test de corrélation paramétrique, on obtient un coefficient de Pearson r = 0,12 qui est associé à une probabilité $p = 0{,}74$, qui est non significative. Ce test permet de tester la pente de la ligne en pointillé sur la *figure 1*. Cette pente n'est pas statistiquement différente de la ligne horizontale (pente égale à 0). Avec le test non paramétrique, on obtient un coefficient de Spearman $r = 0{,}63$ qui est associé à une probabilité $p = 0{,}049$ qui est donc légèrement significative (selon les critères habituellement retenus, $p = 0{,}05$). On est donc dans une situation où deux tests statistiques sur les mêmes données ne produisent pas les mêmes résultats, ce qui conduit à des conclusions différentes. Qui a raison ?

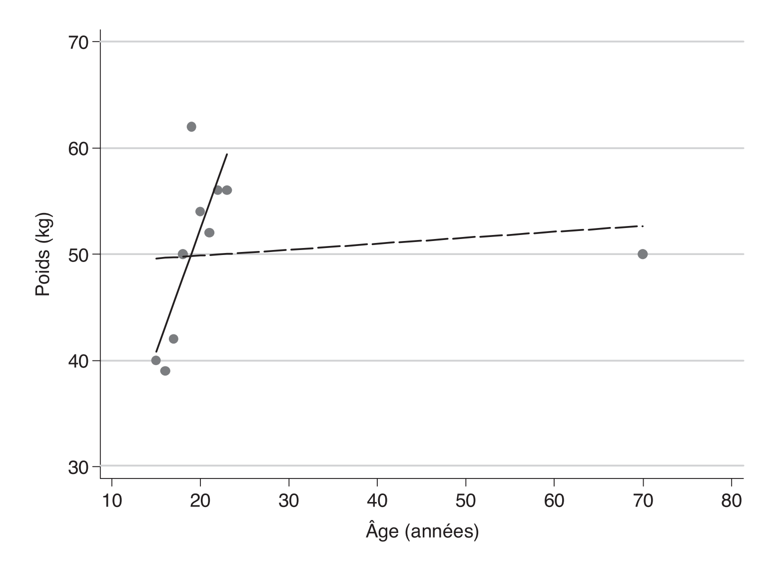


Figure 1. Corrélation entre le poids et l'âge.

En fait, deux questions se posent. Si l'on décide d'inclure dans cette analyse tous les sujets, il vaut mieux réaliser un test non paramétrique qui substitue les rangs aux valeurs brutes : l'influence du sujet âgé de 70 ans devient le même que s'il avait 24 ans, voire 30. Un test non paramétrique n'utilise pas des valeurs réelles, mais simplement leur classement (rang). Cette technique permet d'atténuer l'impact des valeurs extrêmes. En pratique, après avoir trié les données selon l'âge, les sujets jeunes dont l'âge varie entre 15 et 23 ans prennent les places de 1 à 9, tandis que le sujet âgé de 70 ans prend la 10e place : cela rapproche sensiblement sa position de celle des autres.

Pour interpréter ces résultats, on peut se demander ce que fait un sujet de 70 ans dans cette étude. Si on le considère comme une valeur aberrante *(outlier)*, on peut ne pas le prendre en compte. La question devient alors « Y a-t-il une relation entre le poids et l'âge chez les jeunes de moins de 23 ans ? » Par ce simple tour de passe-passe, d'une absence de corrélation (droite en pointillé), on aboutit à une corrélation significative (droite en trait plein).

Dans cet exemple, le nombre de sujets est limité, mais il ne suffit pas d'augmenter la taille de l'échantillon pour limiter l'impact d'une valeur aberrante. Par exemple, en dupliquant les données 4 fois, la corrélation avec le test paramétrique reste faible ($p = 0{,}47$), tandis que le test non paramétrique (à partir des rangs) s'améliore nettement ($p = 0{,}000019$). Le fait de disposer d'effectifs importants n'est donc pas un argument suffisant pour s'orienter vers des tests paramétriques.

Cet exemple concernait deux variables quantitatives. Il en est de même quand une des variables est quantitative et l'autre qualitative (par exemple, le poids et le sexe). Il faut vérifier si la distribution des valeurs quantitatives est conforme à la loi normale dans chaque groupe (ici le poids pour les hommes et les femmes). De surcroît, l'homogénéité des variances des différents groupes doit aussi être contrôlée.

À retenir

Si les distributions sont normales (gaussiennes), on peut utiliser les tests paramétriques.
Si les distributions s'écartent significativement de la courbe de Gauss, il faut utiliser les tests non paramétriques.

Principaux tests statistiques

Tous les tests statistiques sont construits pour évaluer la compatibilité des observations avec l'hypothèse nulle. Quelle que soit la situation, le raisonnement est toujours le même. La statistique qui résume les données est comparée à la valeur de la loi statistique correspondante, c'est-à-dire à la valeur au-dessus de laquelle l'hypothèse nulle doit être rejetée. Le choix de la loi dépend du type de la variable. Avec cette approche, c'est souvent la loi normale qui sert de référence. La plupart des statistiques de test sont conçues de telle sorte que l'on puisse comparer leurs résultats à la distribution gaussienne sous l'hypothèse nulle.

Pour des échantillons de taille inférieure à 30, les tables statistiques de Student doivent être utilisées. Les valeurs seuils varient en fonction de la taille de l'échantillon. Pour des tests non paramétriques, des tables statistiques exactes existent, mais il est également possible d'avoir recours à des approximations.

Dans les sections suivantes, différents cas de comparaison de moyennes, de pourcentages, de médianes et de variances seront considérés.

Comparaison de moyennes

Dans un premier temps, on suppose que la taille des échantillons est suffisante pour garantir des approximations suffisamment précises. La statistique de l'écart réduit Z se calcule en faisant le rapport entre la différence des moyennes et l'écart-type de cette différence. La statistique de test Z suit la loi normale centrée réduite. On vérifie dans la table de la loi normale si l'hypothèse nulle H_0 est acceptable, c'est-à-dire si la valeur de Z est inférieure ou non à la valeur U de la loi normale correspondant à la probabilité recherchée.

$$Z > U_{\alpha/2} \rightarrow \textit{Rejet de } H_0$$
$$Z < U_{\alpha/2} \rightarrow \textit{On ne peut pas rejeter } H_0$$

Par exemple, pour un seuil $\alpha = 0{,}05$, on va comparer la valeur de Z au chiffre $U_{\alpha/2} = 1{,}96$ (on suppose ici que la différence peut être retrouvée dans les deux sens). Pour d'autres valeurs de α, il faut se référer à la table statistique de la loi normale. Plusieurs situations sont décrites ci-dessous. Afin d'éviter les répétitions, les notations suivantes seront utilisées : les groupes seront dénotés par une lettre (A, B, etc.), le nombre de sujets par n, la moyenne par m, son espérance par μ et l'écart-type par s. Une référence à un groupe particulier sera indiquée par un indice du groupe : m_A, s_B…

Un seul groupe (comparaison à une moyenne théorique)

Dans cette situation, on dispose d'un unique échantillon A. La question posée est : la moyenne observée diffère-t-elle significativement d'une valeur connue au préalable, appelée moyenne théorique (μ_0) ? Le choix de cette valeur est très important et correspond à ce que l'on attend suite au résultat observé sur des séries antérieures. On considère le cas bilatéral où est testée l'hypothèse nulle *versus* l'hypothèse alternative :

$$H_0 : \mu = \mu_0 \ \ vs \ \ H_1 : \mu \neq \mu_0.$$

C'est par exemple le cas lorsque l'on désire comparer le taux de réponse à un traitement avec un taux de référence disponible dans la littérature. C'est également le cas quand il s'agit de comparer le pourcentage de toxicité obtenu avec un traitement spécifique aux chiffres des publications précédentes concernant le même traitement. Il peut s'agir aussi de vérifier que la moyenne d'âge des patientes atteintes d'un cancer du sein issues d'une région particulière est analogue à celle des statistiques nationales ou internationales. La statistique de test (centrée-réduite) est calculée par :

$$Z = |m_A - \mu_0| / \left(s_A / \sqrt{n_A}\right).$$

que l'on compare à la valeur $U_{\alpha/2} = 1{,}96$ de la loi normale pour un test bilatéral au niveau de signification de 5 %. L'hypothèse que Z suit la loi normale est une approximation, qui s'avère

généralement suffisante pour des échantillons de grande taille. Dans le cas contraire, deux solutions sont disponibles : soit une autre loi statistique s'avère être mieux adaptée, soit on utilise un test non paramétrique.

Un seul groupe (comparaison de séries appariées)

Dans cette situation, on veut par exemple étudier l'évolution d'un marqueur afin de conclure si oui ou non les moyennes observées avant et après une intervention sur les mêmes sujets sont statistiquement différentes. On dispose donc d'un seul groupe A, mais de deux mesures par sujet (avant et après). L'objectif consiste à comparer la moyenne des différences individuelles à la valeur zéro : cette hypothèse nulle correspond à l'absence de variation de ce marqueur. La statistique de test (centrée-réduite) est calculée par : $Z = m^d_A/s_d$, avec m^d_A la moyenne des différences individuelles de l'échantillon A, qui est calculée sur la différence des valeurs observées (après/avant) pour chaque sujet, et s_d : l'écart-type des différences. La statistique Z est ensuite comparée à la valeur seuil choisie en rapport au degré de signification désiré (généralement $\alpha \leq 0{,}05$).

Deux groupes indépendants

On considère le cas bilatéral où est testée l'hypothèse nulle *versus* l'hypothèse alternative

$$H_0 : \mu_A = \mu_B \text{ vs } H_1 : \mu_A \neq \mu_B.$$

Dans le cas de deux groupes indépendants, l'analyse des résultats doit permettre de conclure si oui ou non les moyennes des valeurs étudiées sont statistiquement différentes entre les deux groupes : par exemple, comparaison au sein de deux groupes de différentes variables cliniques comme l'âge, le poids, l'albumine, etc. La statistique de l'écart réduit Z est calculée par :

$$Z = (m_A - m_B)/s_w,$$

avec s_W étant l'écart-type pondéré des deux échantillons, estimé par

$$s_w = \sqrt{s_A^2/n_A + s_B^2/n_B}.$$

La statistique Z, utilisée pour des échantillons de grande taille, est ensuite comparée à la valeur seuil de la loi normale.

Pour des échantillons de faible effectif (< 30), on peut utiliser le test t de Student ou le test W de Wilcoxon (ou U de Mann-Whitney).

En ce qui concerne le test t de Student, la statistique de test est la suivante :

$$t = (m_A - m_B)/s_c$$

où s_c l'écart-type commun des deux échantillons est estimé par :

$$s_C = \sqrt{\left[(n_A - 1)s_A^2 + (n_B - 1)s_B^2\right] / (n_A + n_B - 2)}.$$

La statistique *t* de Student est ensuite comparée à la valeur seuil, obtenue dans la table de Student à $n_A + n_B - 2$ degrés de liberté (ddl) dans la colonne correspondant à α.

En ce qui concerne le test *W*, ce sont les médianes qui sont comparées. Une fois les observations triées par ordre croissant, la somme des rangs du groupe A est comparée à la valeur seuil de la table correspondant aux effectifs des groupes : ces valeurs seuils ont été obtenues par un échantillonnage exhaustif de la distribution pour un nombre limité de paramètres. Dans le cas de grands échantillons, on peut utiliser l'approximation suivante :

$$W = (R_A - n_A m_r) / (n_A n_B m_r / 6).$$

avec $m_r = (n_A + n_B + 1)/2$ le rang moyen, et R_A la somme des rangs dans le groupe A. Cette statistique centrée-réduite suit une distribution normale.

Comparaison de plusieurs moyennes

Quand il s'agit de comparer les moyennes entre plus de deux groupes, on utilise la statistique de test de Fisher. Le principe de ce test réside dans le raisonnement suivant : si les sujets proviennent de la même population, la variabilité du facteur étudié doit être similaire dans chacun des échantillons. Dans le cadre de trois groupes, l'hypothèse nulle est :

$$H_0 : \mu_A = \mu_B = \mu_C.$$

L'analyse compare la variabilité intragroupe (CM_r, qui mesure dans chaque groupe la différence entre la valeur observée pour chacun des individus et la moyenne du groupe) et la variabilité intergroupes (CM_g, qui mesure la différence entre la moyenne d'un groupe et la moyenne de l'ensemble des groupes). La statistique de test *F* compare la source de variation entre les groupes et la source de variation à l'intérieur des groupes : $F = CM_g / CM_r$, avec

$$CM_g = \frac{SC_g}{K-1} \text{ et } SC_g = \sum_{k=1}^{K}\left(\frac{T_k^2}{n_k}\right) - \frac{T_G^2}{N}$$

et

$$CM_r = \frac{SC_r}{N-k} \text{ et } SC_r = \sum_{k=1}^{K}\sum_{j=1}^{n_k} x_{kj}^2 - \sum_{k=1}^{K}\left(\frac{T_k^2}{n_k}\right),$$

Avec $T_k = \sum_{i=1}^{n_k} x_{ki}$ et $T_G = \sum_{k=1}^{K} T_k$. X_{ki} la valeur de la variable pour le sujet *i* dans le groupe *k* ;

N le nombre total de sujets ; *K* le nombre de groupes ; *k* l'indicateur de groupe, allant de 1 à *K* ; n_k le nombre de sujets dans le groupe *k*. Plus cette variation s'éloigne de un, plus les valeurs

moyennes des groupes diffèrent. Pour appliquer le test statistique, on compare cette valeur de la statistique à la valeur seuil $F_{\alpha}(K - 1, N - K)$ dans la table de Fisher pour un test bilatéral au niveau de signification α, généralement 0,05.

À retenir

L'utilisation du test F à des données manquera de puissance et les conclusions seront erronées si les variances ne sont pas homogènes et les données ne suivent pas une distribution normale. Il n'y a pas trop de sens à comparer des moyennes entre des groupes qui diffèrent par une trop grande hétérogénéité.

Le *tableau I* présente les statistiques descriptives des taux sériques de bêta-2-microglobuline en fonction de l'extension de la maladie de Hodgkin *(figure 2)*.

Tableau I. Statistiques descriptives des taux de bêta-2-microglobuline en fonction du stade.

	I	II	III	IV	Total
N	14	20	25	44	103
Moyenne	2,08	2,54	3,88	3,42	3,177
Écart-type	0,743	1,049	2,765	3,018	2,511
Médiane	2	2,3	2,7	2,45	2,4
Étendu	1-3,5	1,4-5,7	1-11,2	0,4-17,5	0,4-17,5
IQR	0,8	1,3	6.4	5,6	9

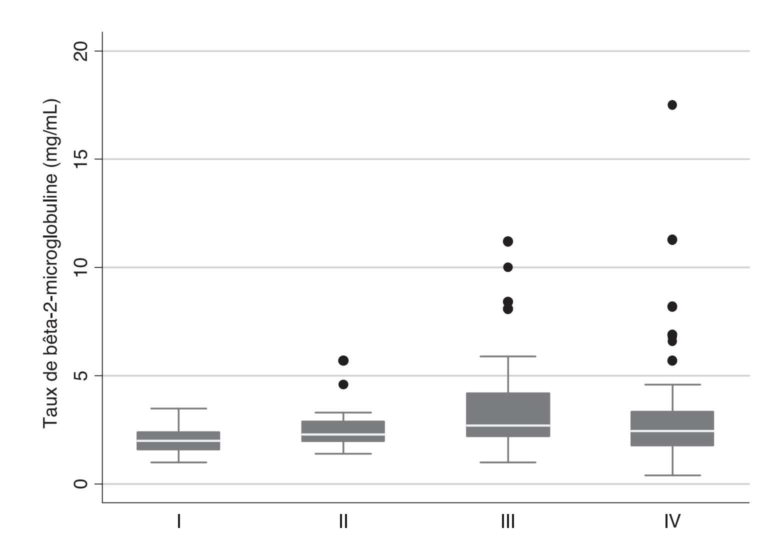


Figure 2. Bêta-2-microglobuline selon le stade d'extension.
La boîte est délimitée par les 25e et 75e percentiles et contient 50 % des observations ; médiane (trait horizontal à l'intérieur de la boîte) ; 5e et 95e percentiles (traits fins horizontaux) ; symboles (valeurs extrêmes).

Le résultat du test $F(3,99) = 2,18$ comparé à la valeur seuil $F_{0,05}(3,99) = 2,70$ est non significatif ($p = 0,095$) mais, avant de conclure, il convient de vérifier les conditions d'application. En effet, on constate que les moyennes augmentent avec les variances en fonction du stade. Ce constat va orienter les analyses statistiques, car ce test suppose la normalité des données et l'égalité des variances entre les groupes. Le test de Shapiro-Wilks (*cf.* page 45) rejette l'hypothèse de normalité et le test de Bartlett (*cf.* page 38) rejette l'hypothèse d'égalité de variances.

Une solution consiste à transformer les variables, par exemple en utilisant une échelle logarithmique *(figure 3)*. Le résultat du test $F_{0,05}(3,99) = 2,25$ est non significatif ($p = 0,087$).

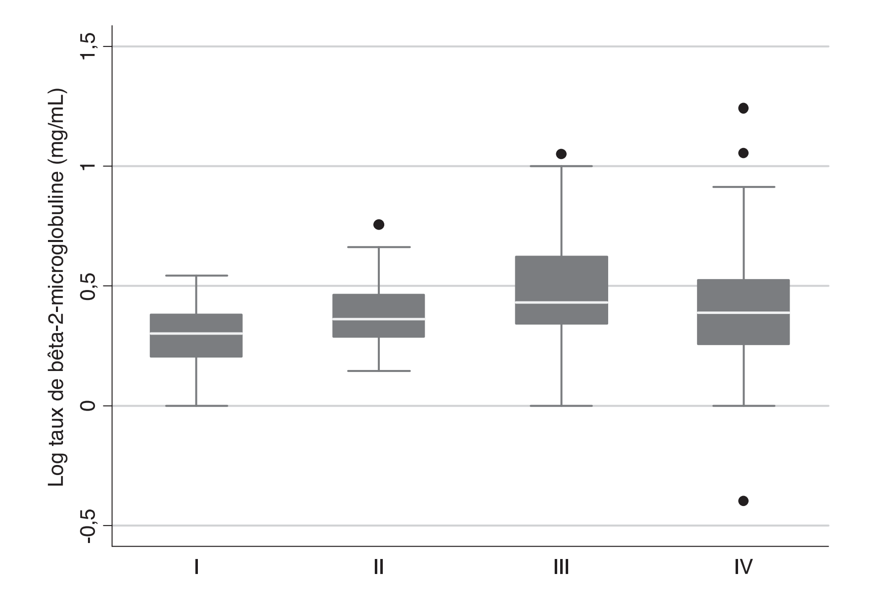


Figure 3. Bêta-2-microglobuline selon le stade d'extension : échelle logarithmique – Boxplot.
La boîte est délimitée par les 25e et 75e percentiles et contient 50 % des observations ; médiane (trait horizontal à l'intérieur de la boîte) ; 5e et 95e percentiles (traits fins horizontaux) ; symboles (valeurs extrêmes).

Le test non paramétrique *H* de Kruskal-Wallis permet de comparer les médianes entre plus de deux groupes. Dans le cas de trois groupes, la statistique *H* se calcule de la manière suivante :

$$H = \frac{12}{N(N+1)}\left[R_A^2/n_A + R_B^2/n_B + R_C^2/n_C\right] - 3(N+1),$$

avec R_k la somme de rangs dans le groupe k et N le nombre total de sujets. Pour plus de trois groupes, la généralisation est facile, car il suffit de rajouter le terme R_D^2/n_D (entre crochets) pour quatre groupes, etc. Pour statuer sur la différence entre les médianes, on recherche la probabilité associée à la statistique H qui suit approximativement une loi de χ^2 à $(K - 1)$ ddl. Le résultat du

test H = 7,80 avec 3 ddl est significatif (p = 0,049). Comme les distributions n'était pas gaussiennes, c'est bien le test H qui doit être préféré même si les variances augmentent malgré la transformation en échelle logarithmique.

Comparaisons multiples post-hoc

Les tests présentés ci-dessus permettent de détecter si oui ou non il y a une différence entre les groupes de manière globale. Pour identifier le sens de la différence observée à l'issue d'une analyse de variance, plusieurs tests peuvent être utilisés, notamment le test de Newman-Keuls, qui est une technique qui compare les moyennes deux à deux après avoir effectué un test de Fisher sur l'ensemble des groupes et dès lors qu'une différence significative a été identifiée. Le test compare la position des moyennes en les classant dans l'ordre croissant : $\mu_1 < \mu_2 < \ldots < \mu_K$, puis en les comparant deux à deux, chacune à leur tour. La première hypothèse teste la plus petite moyenne par rapport à la plus grande :

$$H_0 : \mu_1 = \mu_K \text{ vs } H_1 : \mu_1 \neq \mu_K .$$

Si H_0 est rejetée, c'est-à-dire si une différence significative est notée, on poursuit les comparaisons avec μ_2 et μ_K et ainsi de suite tant que H_0 est rejetée. Au final, on identifie les moyennes qui diffèrent deux à deux. Le test de Newman-Keuls a cependant un inconvénient : il nécessite que les effectifs par sous-groupe soient égaux. Dans la situation où les sous-groupes ont des effectifs inégaux, demeure la possibilité de réaliser autant de tests que de paires de moyennes, mais en changeant le seuil de significativité à l'aide d'une correction de Bonferroni pour comparaison multiple (on divise α par le nombre de comparaisons effectuées).

Comparaison de variances

La comparaison des variances entre deux groupes peut être réalisée par le test paramétrique de Bartlett (1937), par le test non paramétrique d'Ansari-Bradley (Ansari, 1959) ou par le test de Fligner-Killeen (Donnelly, 1999). Ces tests sont relativement complexes et ne seront pas présentés en détail dans ce chapitre. Néanmoins, il est nécessaire de connaître les principes pour ne pas oublier de les utiliser. Les résultats de l'exemple ci-dessus montrent tout leur intérêt.

Test de Fisher-Snedecor

Le test de Fisher-Snedecor est un test paramétrique qui permet de comparer les variances de deux groupes. L'hypothèse nulle est $H_0 : \sigma_A^2 = \sigma_B^2$. La statistique de test F suit une loi de Fisher :

$$F(n_A - 1, n_B - 1) = s_A^2 / s_B^2 .$$

À retenir

Les distributions doivent suivre une loi normale. Le test de Fisher-Snedecor est très sensible aux valeurs extrêmes. Dans ce cas, il est préférable d'utiliser le test d'Ansari-Bradley.

Test de Bartlett

Le test de Bartlett est un test paramétrique qui permet de comparer les variances de plusieurs groupes. L'hypothèse nulle est que les variances sont homogènes. La statistique de test suit une loi de χ^2 à $(K - 1)$ ddl. Legendre *et al.* (2011) ont effectué de nombreuses simulations pour comparer ce type de test, et le test de Bartlett s'est révélé l'un des plus efficaces et robustes, c'est-à-dire qu'il « fonctionne » dans la grande majorité des situations.

Test d'Ansari-Bradley

Le test d'Ansari-Bradley est un test non paramétrique qui permet de comparer la dispersion des valeurs de deux groupes. Dans les domaines de la statistique où les tests paramétriques sont très présents, le test de Bartlett permet de vérifier les hypothèses nécessaires à l'utilisation des tests de Student par exemple. Dans les domaines de la statistique où les tests non paramétriques sont très utilisés, comme en cancérologie, le test d'Ansari-Bradley permet de vérifier les hypothèses nécessaires à l'utilisation des tests de Wilcoxon par exemple. Il ne suffit pas que les distributions ne soient pas normales pour justifier l'utilisation des tests non paramétriques, car la comparaison des médianes nécessite une hypothèse de dispersion semblable entre les groupes.

Test de Fligner-Killeen

Plus souple que le test d'Ansari-Bradley, le test non paramétrique FK permet de comparer les variances de plusieurs groupes. Il est très peu utilisé, mais son utilisation pourrait trouver un essor dans un futur proche car c'est le seul test qui permet de conserver un rapport acceptable entre les erreurs de type I et de type II (Donnelly, 1999).

Comparaison de proportions

Les traitements en cancérologie sont évalués par les bénéfices et les risques (effets positifs et effets négatifs ou délétères) qu'ils peuvent engendrer. Les bénéfices en termes de succès se résument par une proportion de succès. Pour comparer une proportion à une valeur théorique ou pour comparer deux ou plusieurs proportions entre elles, il faut d'abord calculer les proportions, les variances et les écarts-types. Deux approches sont envisagées : l'approche approximative et l'approche exacte.

Comparaison à une proportion théorique

On considère ici un seul groupe A. On veut analyser les résultats afin de conclure si oui ou non le taux de succès observé dans le groupe A est statistiquement différent d'une valeur connue au préalable (proportion théorique π_0). Le choix de cette valeur est très important et doit correspondre au résultat observé sur des séries antérieures similaires.

En suivant une approche semblable à celle utilisée pour la moyenne, la statistique de test (centrée-réduite) est calculée à l'aide de :

$$Z = |p_A - \pi_0|/(s_A),$$

que l'on compare à la valeur $U_{\alpha/2} = 1{,}96$ pour un test bilatéral au niveau de signification $\alpha = 0{,}05$, avec

$$s_A = \sqrt{p_A(1-p_A)/n_A},$$

à condition que les groupes soient constitués d'observations indépendantes.

Une approche plus exacte consiste à calculer l'intervalle de confiance (IC) exact autour de la proportion observée et de situer la proportion théorique par rapport à cet intervalle. Par exemple, on cherche à comparer le taux de succès à 40 %. On observe 7 succès parmi 33 patients, soit un taux de succès de 21 %. L'IC à 95 % varie entre 9 % et 39 %. Observant que la cible 40 % se situe en dehors de l'IC, on peut rejeter l'hypothèse nulle.

Comparaison de proportions (deux groupes)

Notre échantillon est constitué de deux groupes. Il s'agit d'analyser les résultats pour pouvoir conclure si oui ou non les proportions d'un paramètre (réponse au traitement, toxicité, etc.) sont statistiquement différentes dans les deux groupes.

Considérons un exemple où les taux de succès d'un même traitement sont comparés entre deux groupes distincts de 30 patients : 9 (30 %) et 15 (50 %) succès sont observés dans les groupes A et B respectivement.

Comme énoncé précédemment, l'intérêt d'un test statistique porte ici sur la réponse à la question : « est-ce que 30 %, proportion de succès π_A du groupe A, est statistiquement différente de 50 %, proportion de succès π_B du groupe B ? ». Pour cela, on considère le cas bilatéral où est testée l'hypothèse nulle $H_0 : \pi_A = \pi_E$ *versus* $H_1 : \pi_A \neq \pi_B$

Le test utilisé le plus fréquemment dans cette situation (tableau dit de contingences) est le test du chi-2. Comme tout test statistique, il est fondé sur le principe de comparer ce que l'on observe à ce que l'on devrait observer si l'hypothèse nulle était vraie. Sous H_0, on s'attend à ce que la proportion de succès soit le même dans les deux groupes A et B. On suppose donc que les

observations proviennent de la même population, et on calcule le nombre de succès attendu sous cette hypothèse. Ainsi, sur 60 sujets au total, on a observé 9 + 15 = 24 succès (soit 40 %). Sous H_0, on attend donc 12 succès dans chaque groupe, ce qui nous donne un tableau des données sous H_0 :

L'idée du test du chi-2 consiste alors à mesurer l'écart entre les résultats attendus et les résultats observés *(tableau II)* :

$$chi\text{-}2 = \sum_{ij} \frac{(O_{ij} - A_{ij})^2}{A_{ij}}$$

avec i = 1,2 et j = 1,2, les indices associés respectivement aux lignes et aux colonnes. Cela se traduit par le calcul :

$$chi\text{-}2 = \frac{(9-12)^2}{12} + \frac{(15-12)^2}{12} + \frac{(21-18)^2}{18} + \frac{(15-18)^2}{18} = 2{,}5$$

La probabilité p selon la loi du chi-2, associée à la valeur 2,5 (avec 1 ddl) est égale à 0,11 (> 0,05) ; le résultat est donc non significatif (on ne rejette pas H_0, c'est-à-dire que l'on ne rejette pas le fait que les taux de succès soient similaires dans les deux groupes).

Si on observe maintenant les mêmes pourcentages de succès sur 50 sujets dans chaque groupe, la valeur de la statistique de test du chi-2 est égale à 4,16 avec 1 ddl et la valeur de p associée vaut 0,041 (< 0,05). Autrement dit, on rejette l'hypothèse nulle H_0, c'est-à-dire que l'on rejette le fait que les taux de succès soient similaires dans les deux groupes. Pourtant, la question posée est restée la même : « est-ce que 30 % est statistiquement différent de 50 % ? ».

Remarques

Dans le premier cas de figure, on n'affirme pas que le traitement A n'est pas différent du traitement B. Cela veut simplement dire que les observations sur 60 patients ne permettent pas de conclure à une différence entre les deux traitements parce que la différence observée pourrait vraisemblablement être due au hasard. Un effectif plus important pourrait éventuellement lever cette incertitude.

- Les significations clinique et statistique ne doivent pas être confondues. Ainsi, pour une différence d'efficacité assez minime entre deux traitements (de l'ordre de 2 %) et en prenant un effectif important (20 000 personnes, par ex.), nous pouvons « conclure statistiquement » à une plus grande efficacité du traitement A.
- Le degré de liberté correspond à : (nombre de lignes - 1) × (nombre de colonnes - 1). Exemple : pour un tableau de contingences à 2 lignes et 2 colonnes, le degré de liberté est 1 (c'est le cas de notre exemple).
- Pour le test du chi-2, il faut que l'effectif attendu (sous l'hypothèse H_0) de chaque cellule soit supérieur à 5. Ce test utilise une approximation du calcul exact et n'est valable que pour des effectifs de taille suffisante. Dans le cas contraire, il faut utiliser les résultats associés au test de Fisher.

Tableau II. Exemple d'un test du chi-2 avec 60 patients.					
	Groupe A		Groupe B		TOTAL
	Observé	Attendu	Observé	Attendu	
Succès	9	12	15	12	24
Échec	21	18	15	18	36
Total	30	30	30	30	60

Lorsqu'on se trouve dans une situation d'effectifs faibles, le test du chi-2 n'est pas adapté, car il repose sur une distribution asymptotique. Le test exact de Fisher peut alors être utilisé. Supposons le tableau 2 x 2 suivant avec les effectifs observés a, b, c, d.

a	b
c	d

La probabilité *p* que ces effectifs soient distribués de manière aléatoire se calcule avec la formule suivante :

$$p = \frac{(a+b)! \times (c+d)! \times (a+c)! \times (b+d)!}{a! \times b! \times c! \times d! \times (a+b+c+d)!}$$

où (a + b)! signifie factoriel (a + b), c'est-à-dire le produit 2 x 3 x 4... x (a + b)

Comparaison de plus de deux proportions (plusieurs groupes)

Le principe est le même que pour une comparaison de proportions entre deux groupes. La différence repose simplement sur le nombre de groupe à comparer, 3 ou plus. Le test utilisé dans ce cas est également le test du chi-2.

Les hypothèses à vérifier : l'hypothèse nulle consiste à poser que les *K* proportions à comparer ne sont pas différents, tandis que l'hypothèse alternative stipule qu'au moins une des proportions est différente.

Le tableau de contingence ressemble à celui proposé au paragraphe précédent avec bien entendu plus de colonnes et/ou de lignes.

Les contraintes associées à l'utilisation du test du chi-2 sont les mêmes que pour deux groupes. Il faut que les effectifs théoriques du tableau de contingence soient supérieurs ou égaux à 5. Si cette condition n'est pas respectée, il est nécessaire d'envisager des regroupements jusqu'à aboutir, au besoin, à un tableau 2 x 2. Ensuite, les méthodes vues au paragraphe précédent sont applicables.

Test de tendance (proportions sur plusieurs groupes)

Les tests de tendance permettent de déterminer s'il y a une évolution (diminution ou augmentation) d'une proportion (on parle de tendance linéaire). Ces tests ne sont utilisables que si la variable pour laquelle la proportion est calculée est ordinale (année, quantité, niveau, etc.).

Exemple : on observe qu'en 2008 la proportion de patients atteints d'un cancer du côlon ayant le traitement A représente 30 % des patients touchés par ce cancer, en 2009, 34 % et en 2010, 39,2 %. Ici, il s'agit non seulement de déterminer s'il existe une différence de proportion entre ces trois années mais également de définir s'il existe une tendance à la hausse de l'usage du traitement A face à ce type de cancer.

Trois tests permettent d'évaluer une tendance linéaire entre deux variables qualitatives, l'une binaire (présence, absence), l'autre à plusieurs niveaux (année, etc.) : le test de Cochran Armitage, le test de Mantel-Haenszel et le test du rapport de vraisemblance.

Les hypothèses testées sont, pour l'hypothèse nulle, qu'il n'y a pas de différence de proportion et, pour l'hypothèse alternative, qu'il existe une différence et que celle-ci est croissante (ou décroissante).

Nous présenterons ici le test du chi-2 de Cochran-Armitage qui repose sur le calcul de la statistique de test suivant :

$$chi\text{-}2 = \frac{n^3\left[\sum x_i (O_{1i} - A_{1i})\right]^2}{m_1 m_2 \left[n \sum n_i x_i^2 - \left(\sum n_i x_i\right)^2\right]}$$

avec m_1 et m_2 les effectifs prenant respectivement les valeurs 1 et 2 de la variable qualitative binaire, n la population totale de l'échantillon ($m_1 + m_2 = n$), et x représentant la variable ordinale : pour plus de simplicité dans les calculs, il est utile (parfois) de transformer cette variable, par exemple : $x_1 = 0$; $x_2 = 1$... x_n = (dernière – première valeur de la variable ordinale). Une fois la statistique de test calculée, le résultat est comparé à la valeur de la table pour un chi-2 à 1 ddl.

Reprenons notre exemple :

	Années						
	2008 (x_1)		**2009 (x_2)**		**2010 (x_3)**		**Total**
	Observé	**Attendu**	**Observé**	**Attendu**	**Observé**	**Attendu**	
Traitement A	36	42	51	52	55	48	142
Traitement autre	84	78	99	98	85	92	268
Total	120	120	150	150	140	140	410

On modifie les variables x_i : $x'_1 = 0$; $x'_2 = 1$, $x'_3 = 2$

$$chi\text{-}2 = \frac{410 \times [0 \times (36-42) + 1 \times (51-52) + 2 \times (55-48)]^2}{142 \times 268 \left[410 \times (120 \times 0^2 + 150 \times 1^2 + 140 \times 2^2) - (120 \times 0 + 150 \times 1 + 140 \times 2)^2\right]}$$

$\chi = 2{,}88$. il n'y a donc pas de différence significative, aucune tendance significative n'est mise en évidence.

Note : en régression linéaire, l'équation d'une droite est de la forme : $y = a + bx$ et la statistique de test présente ci-dessus revient à tester la pente de la droite de régression *(b)*.

Liaison entre deux variables quantitatives

Dire que deux variables quantitatives *X* et *Y* sont corrélées, c'est dire qu'il existe une liaison statistique entre ces deux variables. Si la corrélation mise en évidence est positive avec *Y* qui augmente quand *X* augmente, on dira que les deux variables varient dans le même sens ; si la corrélation est négative, on dira que les deux variables varient en sens opposé.

Pour évaluer l'intensité de la liaison entre deux variables quantitatives prenant des valeurs numériques (entières ou décimales), indépendantes d'un sujet à un autre, on peut utiliser soit la méthode paramétrique de Pearson, soit la méthode non paramétrique de Spearman, qui permettent toutes deux de calculer un coefficient de corrélation. Par définition, on dispose de deux mesures sur le même individu, par exemple le dosage de deux marqueurs *X* et *Y*.

Le coefficient de corrélation varie entre -1 (la corrélation est négative) et +1 (la corrélation est positive) ; lorsque celui-ci est proche de 0, il n'y a pas de corrélation. La valeur absolue du coefficient de corrélation indique l'intensité de la liaison.

On pose deux hypothèses : l'hypothèse nulle H_0 d'absence de liaison entre les deux variables *X* et *Y*, contre l'hypothèse alternative H_1 d'existence d'une liaison. Sous H_0, le coefficient de corrélation doit être proche de zéro. Pour estimer la validité de chacune des hypothèses, la valeur absolue du coefficient de corrélation est comparée à 0. Sous H_0, le rapport de ce coefficient sur son écart-type suit une loi de Student à *N-2* degrés de liberté.

Coefficient r de Pearson

Les conditions d'utilisation de la méthode de Pearson impliquent que les distributions des variables suivent une loi normale. Le coefficient de corrélation *r* de Pearson se calcule de la manière suivante :

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]\left[\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right]}}$$

Les valeurs x_i et y_i correspondent aux valeurs brutes des observations pour l'individu *i*. La statistique de test $t_r = r / s_r$ avec l'écart-type estimé par

$$s_r = \sqrt{(1 - r^2)/(n - 2)}$$

est ensuite comparée à la valeur seuil obtenue dans la table statistique de Student.

Coefficient r de Spearman

La méthode de Spearman est fondée sur une approche non paramétrique qui compare le rang des valeurs de chacune des deux variables, et non pas leurs valeurs elles-mêmes. Le coefficient de corrélation r de Spearman se calcule de la manière suivante :

$$r_s = 1 - 6\sum(x_i - y_i)^2 / n(n^2 - 1)$$

Les valeurs x_i et y_i correspondent ici aux rangs des observations pour l'individu *i* et non à leurs valeurs brutes. Dans le cas extrême où les données sont parfaitement rangées avec les plus petites valeurs de X correspondant aux plus petites valeurs de Y, on voit que leur différence est de zéro et donc $r_s = 1$.

La statistique de test $t_{rs} = r_s / s_{rs}$ avec l'écart-type estimé par

$$s_{r_s} = \sqrt{(1 - r_s^2)/(n - 2)}$$

est, à l'instar du coefficient de Pearson, comparée à la valeur seuil obtenue dans la table statistique de Student. Une autre possibilité consiste en l'utilisation d'un test de permutation.

Les tests de normalité

Nous ne décrirons que deux tests de normalité qui permettent de prendre en charge la très grande majorité des cas. Celui de Kolmogorov-Smirnov qui peut être adapté à tout test de l'adéquation d'une distribution avec une loi quelconque (de Gauss, de Poisson, etc.) et celui de Shapiro et Wilks plus adapté à notre avis aux petits échantillons que le premier. Ces deux tests seront présentés et appliqués à l'exemple suivant : soit 10 valeurs ordonnées : 23, 23, 24, 24, 25, 25, 60, 61, 63, 63, dont la répartition bimodale est évidente. La moyenne est estimée à 39,1, l'écart-type à 19,5, la médiane à 25 et l'intervalle interquartile à [24 ; 60] d'étendue 60 - 24 = 36.

Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov permet de vérifier la normalité de la variable (quantitative) d'intérêt. On compare l'écart entre la fonction de répartition de la variable et la fonction de répartition de la loi normale. Une fonction de répartition F d'une variable aléatoire X se définit pour tout

réel x par : $F(x) = P(X \leq x)$. $F(x)$ est la probabilité que la variable aléatoire X prenne une valeur inférieure ou égale à x. L'hypothèse H_0 suppose que la distribution suit la loi normale. C'est en calculant la différence maximale sur l'échantillon entre la répartition observée et celle de la loi normale que le test conclut.

Après centrage et réduction sur la moyenne et la variance observées, ces valeurs deviennent : - 0,825, - 0,825, - 0,773, - 0,773, - 0,722, - 0,722, + 1,070, + 1,122, + 1,224, + 1,224. Si ces données suivent une distribution normale, les probabilités d'être inférieur à chacune de ces valeurs sont : 0,204, 0,204, 0,220, 0,220, 0,235, 0,235, 0,858, 0,869, 0,890, 0,890 tandis que dans notre échantillon, la probabilité d'être inférieur à chaque valeur correspond à la fréquence cumulée. L'écart maximal (0,365), qui correspond à la valeur 25, est à la limite de la signification ($p = 0,07$).

Shapiro et Wilks

Ce test est réputé particulièrement efficace quand la distribution des données est asymétrique. Nous en conseillons l'utilisation quand l'effectif est compris entre 4 et 2 000. Ce test repose sur deux estimateurs de la variance de l'échantillon :

- le premier correspond aux étendues partielles : $x_n - x_1$, $x_{n-1} - x_2$... ;
- le second est la somme des carrés des écarts à la moyenne : $SCE = n\,S^2$.

où n est le cardinal de l'échantillon, S l'écart-type et $x_1, x_2, \ldots x_n$, les valeurs triées.

Si n est pair, le résultat à calculer est le rapport :

$$W = \frac{\sum_{i=1}^{n/2} (a_i d_i)^2}{SCE}$$

Et si n est impair :

$$W = \frac{\sum_{i=1}^{(n-1)/2} (a_i d_i)^2}{SCE}$$

avec a_i les coefficients donnés dans une table fonction de n et de i. Ce résultat est à comparer à la valeur seuil fournie dans les tables. Si $W < W_{1-\alpha,n}$ alors la distribution n'est pas considérée comme gaussienne. Pour l'exemple présenté dans le paragraphe « Comparaison de plusieurs moyennes » supra, $W = 0,683$ et $W_{1-\alpha,n} < 0,781$, valeur donnée dans la table, colonne 1 %, ligne 10. On peut raisonnablement rejeter l'hypothèse H_0 de normalité.

Références

- Shapiro S, Wilks M. An analysis of variance test for normality (complete samples). *Biometrika* 1965 ; 53 : 591-611.
- Spearman C. The proof and measurement of association between two things. *Am J Psychol* 1904 ; 15 : 72-101.
- Siegel S. *Non parametric statistics*. New-York : McGraw-Hill, 1956, 312 pages.
- Bartlett MS. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London Series A* 1937 ; 160 : 268-82.
- Hollander M, Wolfe DA. *Nonparametric Statistical Methods*. New York : John Wiley & Sons, 1973 : pages 83-92.
- Conover WJ, Johnson ME, Johnson MM. A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 1981 ; 23 : 351-61.
- Legendre P, Borcard D. Statistical comparison of univariate tests of homogeneity of variances. *J Stat Comput Simul* 2011 (soumis).
- Bouyer J. *Méthodes statistiques Médecine-Biologie*. Paris : ESTEM, les éditions INSERM, 1996, 353 pages.
- Ansari AR, Bradley RA. Rank-Sum for dispersions. *The Annals of Mathematical Statistics* 1960 ; 31 (4) : 1174-89.
- Donnelly SM, Kramer A. Testing for multiple species in fossil samples: An evaluation and comparison of tests for equal relative variation. *Am J Phys Anthropol* 1999 ; 108 (4) : 507-29.
- Armitage P. Tests for linear trends in proportions and frequencies. *Biometrics* 1955 ; 11 : 375-86.
- Liu H. Cochran-Armitage trend test using SAS, Paper SP05, Merck Research Labs, Merck & Co., Inc, Rahway, NJ, 2007.
- Bernard PM. *Analyse des tableaux de contingence en épidémiologie*. Québec : Presses de l'Université du Québec, 2004.

Quand dit-on qu'une différence est statistiquement significative ?

C. Hill, A. Laplanche

Quand dit-on qu'une différence est statistiquement significative ? Si vous n'avez jamais su la réponse à cette question, vous devez être très agacé de l'importance que cette formule magique a acquise dans la littérature médicale. L'objectif de ce chapitre est de montrer, à partir d'un cas particulier simple, ce que cela veut vraiment dire. En perdant son mystère, le fameux *p*, qui accompagne en général la conclusion de différence significative, devrait acquérir plus d'intérêt et représenter un résultat numérique aussi facile à discuter qu'une réduction de l'hypertension de 10 mm Hg ou une survie à 5 ans passant de 50 % à 57 %.

Calcul de *p* dans une situation simple

L'exemple très simple est le suivant : deux groupes égaux de patients ont été constitués par tirage au sort. Un groupe a été traité par un traitement A, l'autre par un traitement B. Au total, 10 décès ont été observés. Ces 10 décès se répartissent dans les deux groupes de traitement de la façon suivante : 8 se sont produits dans le groupe traité par A et 2 dans le groupe traité par B. La question est de savoir si l'un des traitements est meilleur que l'autre, c'est-à-dire si la mortalité est vraiment différente entre les deux groupes. L'autre possibilité est que les deux traitements soient en réalité d'efficacité identique, la différence observée étant alors l'effet du hasard seul.

Pour savoir si la différence est « statistiquement significative », il nous faut étudier ce qui peut arriver quand les traitements sont équivalents.

L'inventaire des possibilités est facile à faire : si l'on observe 10 décès au total, ceux-ci peuvent se répartir dans les deux groupes de 11 manières différentes depuis 0 dans le groupe A (et donc 10 dans le groupe B), jusqu'à 10 dans le groupe A (et donc 0 dans le groupe B). Ces possibilités sont listées dans le *tableau I*.

Tableau I. Répartition possible de 10 échecs observés dans deux groupes de traitement.			
Nombre d'échecs		**Nombre de tirages possibles**	**Probabilité**
Groupe A	**Groupe B**		
0	**10**	1	**0,000977**
1	**9**	10	**0,00977**
2	**8**	45	**0,0439**
3	7	120	0,117
4	6	210	0,205
5	5	252	0,246
6	4	210	0,205
7	3	120	0,117
8	**2**	45	**0,0439**
9	**1**	10	**0,00977**
10	**0**	1	**0,000977**

Nous avons besoin de connaître la probabilité de chacune de ces possibilités quand les traitements sont équivalents. Cela est un exercice que vous avez tous fait au lycée dans un autre contexte. Le problème est en effet équivalent à celui du tirage de 10 boules dans une boîte (une urne dans le jargon des probabilités) qui contient moitié de boules blanches et moitié de boules noires. Le nombre de boules est si grand que le fait d'avoir retiré une boule blanche, par exemple, ne change pas la probabilité que la seconde soit blanche qui reste donc égale à 1/2. La probabilité d'observer 10 boules blanches sur 10 ou bien les 10 échecs dans le groupe B est égale à :
$1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2 \times 1/2$, que l'on peut aussi écrire : $(1/2)^{10}$ et qui vaut 1/1024 ou encore : 0,000977.

Reportons cette probabilité dans la colonne « probabilité » du *tableau I*. Le calcul est le même pour 10 boules noires, nous avons donc rempli deux lignes de cette colonne. Pour les autres situations, les calculs sont un tout petit peu plus compliqués par le fait qu'il y a plusieurs façons de tirer, par exemple, une boule noire et 9 blanches. Chaque façon a la même probabilité 1/1024 et correspond à un ordre de tirage défini : par exemple, une première boule noire et les 9 autres blanches ou une première boule blanche, une seconde noire et les 8 autres blanches, etc. Le nombre de tirages possibles, ici 10, est indiqué dans le *tableau I*. La probabilité d'observer une noire et 9 blanches est égale au nombre de tirages possibles multiplié par 1/1024 : soit 10/1024 ou encore 0,0097.

À ce stade, si cette histoire a réveillé en vous les plaisirs oubliés de la combinatoire, si les notations *n!* ou : $C_n^p = \binom{n}{p}$ vous disent encore quelque chose, vous devriez pouvoir reconstituer l'ensemble de la colonne du nombre des tirages possibles et donc des probabilités.

Nous nous sommes apparemment beaucoup éloignés de notre problème de départ, dans lequel nous avions observé 8 décès dans un groupe traité par A et 2 dans un groupe de même taille traité par B. En réalité, nous avons pratiquement résolu le problème. En effet, le *p* que nous cherchons est la probabilité de trouver une différence au moins aussi grande que celle observée, simplement par hasard, c'est-à-dire si les deux traitements sont équivalents. Cette probabilité peut se lire presque directement dans le *tableau I* : en effet, les situations correspondant à une différence au moins aussi grande que 8 décès dans un groupe et 2 dans l'autre sont indiquées en caractères gras et *p* est la somme des probabilités de ces situations au moins aussi extrêmes, soit :

p = Prob (8 & 2 ou 9 & 1 ou 10 & 0 ou 2 & 8 ou 1 & 9 ou 0 & 10)

$p = 0{,}0439 + 0{,}00977 + 0{,}000977 + 0{,}0439 + 0{,}00977 + 0{,}000977 = 0{,}109$

La différence est-elle statistiquement significative ?

La différence n'est pas significative puisque le *p* observé est supérieur à 0,05 (5 %) ; en effet, pour des raisons purement historiques – on raconte que le statisticien britannique Ronald Fisher, qui voulait publier des tables, a été obligé de choisir une limite, les tableaux plus complets étant protégés par un copyright – et aucunement scientifiques, on convient d'appeler différence statistiquement significative toute différence qui a moins de 5 chances sur 100 de se produire simplement par hasard. Cette convention a, malheureusement, pris une importance beaucoup trop grande dans la mesure où la formule magique « différence significative » conduit souvent à oublier de regarder la valeur de *p*. En raisonnant de façon très (trop) simpliste, c'est-à-dire en ignorant tout le reste des informations (pharmacologiques, sur l'animal, etc.), un essai thérapeutique étudiant un tout nouveau produit et dont les résultats sont tout juste significatifs ne constitue pas une expérience suffisamment convaincante pour conclure à l'efficacité d'un médicament et le mettre sur le marché. En effet, on mettrait ainsi sur le marché environ 5 % des placebos ou des produits inactifs étudiés. L'exemple choisi permet aussi d'expliquer la différence entre test unilatéral et test bilatéral. Cette différence n'est pas essentielle au premier abord et le sujet est donc traité en annexe.

Maintenant que la formule « la différence est significative » n'est plus magique, comment peut-on interpréter une valeur de *p* ? Quelle est la valeur de *p* qui doit emporter notre conviction ? Si *p* = 5 % n'est pas suffisant pour mettre une nouvelle molécule sur le marché, faut-il un *p* de 1 pour 1 000 ? De 1 pour 10 000 ? Il n'y a pas de réponse unique à cette question, cela dépend de l'ensemble des données disponibles sur le problème. Si un premier essai d'une molécule peu toxique montre une augmentation significative de la survie de patients atteints d'une maladie jusqu'ici toujours mortelle, avec un *p* égal à 5 %, ce résultat peut être considéré comme très prometteur. Si un essai montre qu'un produit homéopathique est supérieur à un placebo avec

p = 5 %, on pourra penser que cet essai a été publié parce qu'il montrait une différence significative, alors que 19 autres essais négatifs sont restés dans les placards des mêmes investigateurs ou d'autres investigateurs. Le *p* est donc un élément à prendre en compte parmi d'autres. C'est précisément pour cela qu'une bonne compréhension de ce qu'il représente est nécessaire.

Différence entre test unilatéral et test bilatéral

Nous allons expliquer la différence entre test unilatéral et bilatéral (*one-sided* et *two-sided tests*), et la relation avec *p*. Dans notre exemple, nous nous sommes intéressés à une situation bilatérale, c'est-à-dire que nous nous sommes demandés si les traitements étaient différents, sans idée *a priori* sur le sens de cette différence. On peut supposer que le traitement A est un traitement de référence (ou un placebo) et B un nouveau produit et calculer seulement la probabilité d'observer une différence au moins aussi grande que celle observée, les seules possibilités à envisager étant celles dans lesquelles B est supérieur à A ; cela constitue un test unilatéral. On a alors :

p = Prob (8 & 2 ou 9 & 1 ou 10 & 0)

$p = 0{,}0439 + 0{,}00977 + 0{,}000977 = 0{,}055$

La différence est maintenant à la limite de la signification. Cet exemple illustre bien le caractère arbitraire de la limite de 5 %, et l'importance capitale de l'information sur la nature uni- ou bilatérale du test.

Nous pensons personnellement que les tests présentés devraient être, en règle générale, toujours bilatéraux ; l'expérience prouve qu'il arrive qu'un nouveau traitement soit significativement pire qu'un placebo. On voit parfois utiliser la notation 2p pour désigner les valeurs de *p* correspondant à des tests bilatéraux, cela est une façon simple et rapide de préciser que le test correspondant est bilatéral.

À retenir

La valeur de *p* est la probabilité que la différence soit au moins aussi grande que celle qui a été observée, simplement par hasard.
Par convention, on déclare significatives les valeurs de *p* inférieures à 5 %.

Statistiques bayésiennes

P. Roy, R. Porcher

Les données d'une étude, d'un essai clinique n'ont pas d'intérêt en soit. Ainsi, l'objectif d'un essai comparatif randomisé de phase III n'est pas de comparer l'efficacité de deux traitements sur deux groupes de patients aléatoirement désignés pour recevoir les traitements A ou B, mais d'utiliser les résultats de cet essai pour appréhender la vraie différence d'efficacité des traitements comparés, par exemple la différence entre les probabilités de guérison associées aux traitements (respectivement π_A et π_B). L'interprétation des résultats d'une étude dépasse donc la seule prise en compte de ses résultats puisqu'il s'agit de répondre à une question plus large, plus générale. En quoi les résultats de cette étude contribuent-ils à améliorer notre connaissance de la différence des probabilités de guérison des traitements comparés ? Cette question est celle de l'inférence statistique et plus particulièrement de l'estimation d'un paramètre d'intérêt, comme la différence entre les probabilités de guérison associées aux deux traitements : $\theta = \pi_A - \pi_B$. Les approches fréquentiste et bayésienne représentent des apports complémentaires dans le domaine de la recherche clinique en cancérologie. Dans le domaine de l'essai clinique, l'approche bayésienne apparaît bien adaptée à l'accumulation progressive d'informations sur l'efficacité et la toxicité de traitements. Ce chapitre vise à donner dans un premier temps quelques éléments comparatifs des approches fréquentistes et bayésiennes, puis à présenter quelques applications des approches bayésiennes pour la recherche clinique en cancérologie.

Inférences fréquentistes et inférences bayésiennes

L'un des éléments permettant de distinguer les statistiques fréquentistes et bayésiennes est la direction de l'inférence. L'approche fréquentiste est déductive. Elle part de l'hypothèse et la confronte aux données. L'approche bayésienne est inductive. Elle part des données pour estimer la distribution du paramètre inconnu [1].

L'approche fréquentiste

L'approche fréquentiste, déductive, s'appuie sur la théorie de l'urne, *e.g.* les propriétés des distributions d'échantillonnage. Le paramètre inconnu est fixe, l'analyse s'appuie sur la distribution d'échantillonnage de son estimateur. Ainsi, pour démontrer une différence entre les probabilités

de guérison des traitements A et B, deux hypothèses sont préalablement spécifiées. L'hypothèse nulle est celle de probabilités de guérison identiques $H_0 : \theta = 0$. L'hypothèse alternative est, dans le cas d'une formulation bilatérale, celle d'une différence d'efficacité $H_1 : \theta \neq 0$. Une variable aléatoire « grandeur test » Z est alors utilisée, rapportant l'écart entre les estimateurs des probabilités de guérison à son erreur-type, sous l'hypothèse nulle H_0. La distribution de Z sous l'hypothèse nulle H_0 découle directement du schéma de l'urne. Dans l'exemple utilisé, sous certaines conditions d'effectif qu'il n'est pas utile de rappeler ici, Z suit une loi normale centrée réduite :

$$Z = \frac{\hat{\pi}_A - \hat{\pi}_B}{\sqrt{\hat{\pi}_0(1-\hat{\pi}_0)\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}} \sim N(0,1)$$

$\hat{\pi}_0$ étant l'estimateur de la probabilité de guérison commune sous l'hypothèse nulle, n_A et n_B les effectifs des groupes de patients respectivement traités par A ou B. Cette variable aléatoire prend une valeur z à partir des données d'une étude, π_A, π_B, π_C étant estimés à partir de p_A et p_B, les proportions de patients guéris dans les deux groupes. Le petit *p*, probabilité d'observer un résultat au moins aussi éloigné de l'hypothèse nulle que celui observé dans l'étude courante, se déduit de la distribution de Z. Des valeurs faibles de *p*, inférieures au risque de première espèce α, conduisent au rejet de l'hypothèse nulle. Cette probabilité *p* est parfois interprétée à tort comme la probabilité de l'hypothèse nulle alors qu'il s'agit de la probabilité d'observer, sous H_0, un résultat au moins aussi éloigné de H_0 que le résultat courant. Le rejet de l'hypothèse nulle conduirait ici à accepter une hypothèse alternative de différence d'efficacité. Il existe par définition une infinité de valeurs de θ vérifiant $H_1 : \theta \neq 0$. Le risque de seconde espèce *β* n'est défini que lorsqu'une hypothèse alternative particulière est spécifiée *a priori*. La puissance, complément à 1 de *β*, est la probabilité de rejeter l'hypothèse nulle sous l'hypothèse alternative spécifiée. Le nombre de sujets nécessaires est l'effectif permettant, sous l'hypothèse alternative spécifiée, de rejeter l'hypothèse nulle avec la puissance désirée. Par définition, il n'y a pas de sens à calculer la puissance *a posteriori* [2]. Ainsi, rejeter l'hypothèse nulle ne signifie aucunement accepter l'hypothèse alternative spécifiée pour le calcul du nombre de sujets nécessaires. L'estimation (fréquentiste) ponctuelle de la différence d'efficacité est la valeur prise par la variable aléatoire θ sur l'étude courante : $p_A - p_B$. La définition de l'intervalle de confiance au risque α de la différence d'efficacité

$$(p_A - p_B) \pm u_{\alpha/2}\sqrt{\frac{p_A(1-p_A)}{n_A} + \frac{p_B(1-p_B)}{n_B}}$$

découle également du schéma de l'urne ($u_{\alpha/2}$ est la valeur associée par la fonction de répartition de loi normale centrée réduite à la probabilité α/2). Si l'étude était reproduite un très grand nombre de fois, (1 - α) des intervalles de confiance obtenus contiendraient la vraie valeur de *θ*. Cette approche déductive se retrouve dans la définition de la vraisemblance. La vraisemblance de la valeur du paramètre *θ*, $p(y|\theta)$, est la probabilité des données si le paramètre avait cette valeur, la barre verticale soulignant la nature conditionnelle de cette probabilité. C'est donc une quantité

qui mesure « l'accord » des données avec une valeur du paramètre. Elle nous renseigne sur les différentes valeurs du paramètre compatibles avec les données. La méthode du maximum de vraisemblance consiste à retenir comme estimation du paramètre la valeur de θ qui maximise cette probabilité. C'est une méthode générale de construction d'estimateurs qui s'étend naturellement à un ensemble de paramètres. La vraisemblance d'un modèle est la vraisemblance des estimations du maximum de vraisemblance de ses paramètres.

L'approche bayésienne

L'approche bayésienne, inductive, vise à estimer la distribution de probabilité du paramètre d'intérêt $p(\theta|y)$ à partir des données analysées. La connaissance que l'on a du paramètre d'intérêt θ est donc estimée par sa loi de probabilité *a posteriori*. De l'application du théorème de Bayes

$$p(\theta|y) = \frac{p(y|\theta)}{p(y)} \times p(\theta)$$

se déduit la relation de proportionnalité entre les termes comprenant θ

$$p(\theta|y) \propto p(y|\theta) \times p(\theta).$$

où ∝ est le symbole de proportionnalité. Connaissant les données, la distribution de la probabilité *a posteriori* du paramètre d'intérêt $p(\theta|y)$ est donc un compromis de la distribution *a priori* $p(\theta)$ qui représente la connaissance du paramètre avant l'expérience et de ce que nous apprennent les données sur θ : la vraisemblance $p(y|\theta)$. La distribution *a posteriori* représente le niveau d'incertitude ou de connaissance que l'on a du paramètre θ une fois les données analysées. Selon la forme de la distribution de $p(\theta|y)$, la moyenne, la médiane ou le mode de cette distribution peuvent être retenus comme estimations ponctuelles bayésiennes du paramètre θ. Un intervalle, uni ou bilatéral, ayant une probabilité fixée de contenir θ, par exemple 95 %, peut être construit à partir de la distribution de $p(\theta|y)$. La définition de cet intervalle, appelé intervalle de crédibilité, diffère de celle de l'intervalle de confiance fréquentiste. L'intervalle de crédibilité à 95 % est ici l'intervalle contenant 95 % des valeurs possibles du paramètre θ dont la distribution de probabilité *a posteriori* a été estimée. Des intervalles de crédibilité peuvent également être construits à partir des valeurs les plus élevées de la densité de $p(\theta|y)$. Pour une même probabilité de contenir θ, les intervalles obtenus sont alors plus étroits en cas de distributions asymétriques de $p(\theta|y)$, et peuvent regrouper plusieurs intervalles disjoints en cas de distribution multimodale de $p(\theta|y)$ [3]. Lorsque la distribution *a priori* $p(\theta)$ est uniforme (distribution *a priori* non informative), la probabilité *a posteriori* que le paramètre d'intérêt soit compris dans l'intervalle de confiance fréquentiste vaut évidemment $(1 - \alpha)$. La distinction entre les intervalles de confiance et de crédibilité peut alors apparaître excessive, mais nous avons vu qu'il s'agit de concepts différents.

La distribution *a posteriori* découle de la combinaison de la distribution *a priori* $p(\theta)$ et de l'apport des données expérimentales $p(y|\theta)$, leurs contributions respectives dépendant de l'information *a priori*

disponible sur θ et de celle apportée par les données. La distribution *a priori* résume la connaissance disponible sur le paramètre d'intérêt. Il ne s'agit pas d'un *a priori* chronologique, mais de la prise en compte de l'information sur θ extérieure aux données de l'étude courante, en l'absence des données de l'étude. L'information apportée par les données d'une étude de petite taille ou retrouvant une ampleur d'effet modérée pourra bénéficier de la prise en compte d'une distribution *a priori* informative, celle apportée par les données d'une étude de grande taille estimant une ampleur d'effet conséquente influera davantage sur la distribution de $p(\theta|y)$ [3, 4]. Lorsque la distribution *a priori* est non informative, donnant des probabilités identiques pour les différentes valeurs de θ, *e.g.* une distribution de $p(\theta)$ uniforme dans les cas continu, les résultats ne dépendent que des seules données. La prise en compte de l'information *a priori* revient à ajouter aux données de l'étude des « observations fictives » dont le nombre dépend du niveau d'incertitude que l'on a sur θ. Utiliser une distribution *a priori* informative revient à calibrer une étude fictive fournissant la même information sur θ que celle apportée par la distribution de $p(\theta)$, les données de l'étude courante contribuant à améliorer la précision de l'estimation bayésienne [3]. La combinaison d'informations issues de plusieurs sources diminue progressivement l'incertitude sur le paramètre d'intérêt. Le choix de la distribution *a priori* est un élément important de la démarche bayésienne. Un *a priori* sceptique remettra en doute l'efficacité du traitement et contribuera à déclarer les résultats de l'étude « trop bons ». Un *a priori* enthousiaste renforcera la conviction d'un résultat positif. Le choix de la distribution *a priori* peut s'appuyer sur les données antérieures disponibles. La réalisation d'une analyse de sensibilité (ou analyse de robustesse), visant à étudier la répercussion de la distribution *a priori* sur le résultat final, est une étape indispensable de l'approche bayésienne.

Dans l'approche bayésienne, le choix entre deux hypothèses H_0 et H_1 complémentaires ($p(H_1) = 1 - p(H_0)$) s'appuie sur la comparaison de leurs probabilités *a posteriori*. En appliquant le théorème de Bayes, le rapport de leurs probabilités *a posteriori* s'écrit :

$$\frac{p(H_0|y)}{p(H_1|y)} = \frac{p(y|H_0)}{p(y|H_1)} \times \frac{p(H_0)}{p(H_1)}.$$

Ainsi l'odds des probabilités *a posteriori* est obtenu en multipliant l'odds des probabilités *a priori* par le rapport de vraisemblance (ou facteur de Bayes), dont les valeurs sont comprises entre 0 et l'infini. Dans l'approche fréquentiste, le test statistique est le test de l'hypothèse nulle, des valeurs de p inférieures au risque de première espèce conduisant à rejeter l'hypothèse nulle. L'approche bayésienne ne privilégie pas H_0 ou H_1. L'étude de leurs probabilités peut conduire à retenir l'une de ces hypothèses.

L'approche bayésienne permet d'estimer la distribution prédictive *a posteriori* d'une valeur future de y (appelé *y-tilda*), qui serait obtenue, par exemple, après l'inclusion de nouveaux sujets dans l'étude.

$$p(\tilde{y}|y) = \int p(\tilde{y}|y, \theta) p(\theta|y) d\theta.$$

La distribution *a posteriori* $p(\theta|y)$ estime notre incertitude courante sur θ. Sous l'hypothèse que, conditionnellement à θ, les valeurs de *y-tilda* et *y* sont indépendantes

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta$$

La distribution de *y-tilda* découle de l'intégration sur les connaissances actuelles de θ. Les distributions bayésienne prédictives *a posteriori* sont utilisées pour la planification d'études, le calcul de puissance, la validation de modèles, les analyses séquentielles et, en particulier, l'application de règles d'arrêt et l'analyse de la décision médicale.

Application de l'approche bayésienne

Prise en compte des utilités et de l'information *a priori* pour la planification des essais

Les essais comparatifs prévoient classiquement une répartition équilibrée des sujets entre les bras test et contrôle pour maximiser la puissance des tests réalisés. Berry [5] propose d'optimiser les schémas d'études en maximisant le gain thérapeutique attendu sur l'ensemble des patients qui participent à l'essai et bénéficieront du traitement après l'essai. La taille de la population cible et la distribution *a priori* des effets thérapeutiques attendus dans les bras test et contrôle influencent alors le nombre de sujets à inclure et leur répartition entre les bras test et contrôle.

Modélisation hiérarchique bayésienne, méta-analyses

L'approche bayésienne permet de réduire l'incertitude sur le paramètre d'intérêt θ en combinant les informations de plusieurs sources. La modélisation hiérarchique s'appuie sur un modèle à effets aléatoires hiérarchique à deux niveaux. Le premier niveau prend en compte l'hétérogénéité entre études et le second l'effet patient emboîté dans l'étude. L'analyse hiérarchique bayésienne fournit une estimation de la distribution *a posteriori* du paramètre θ pour l'ensemble de la population des études et des estimations des distributions des probabilités *a posteriori* des θ_j de chaque étude. Ces dernières dépendent de l'information issue des données de l'étude concernée et de l'information issue des données de l'ensemble des études. Elles correspondent aux valeurs prédites pour de nouvelles observations. L'hétérogénéité entre études varie théoriquement entre deux extrêmes. En l'absence d'hétérogénéité, chaque étude est un échantillon contribuant à l'estimation de la distribution *a posteriori* d'un paramètre unique θ. Les données des différentes études peuvent alors être regroupées pour cette estimation [5]. À l'opposé, en présence d'études totalement indépendantes, les données d'une étude n'apportent aucune information pour estimer la distribution *a posteriori* des paramètres des autres. Entre ces deux extrêmes se trouve la situation plus réaliste de paramètres ni identiques ni indépendants mais interchangeables, c'est-à-dire similaires dans le sens qu'il n'y a pas de raison *a priori* pour que l'un d'entre eux soit systématiquement différent des autres. Cela revient à considérer que les θ_j ont été échantillonnés dans une distribution aléatoire [3, 4]. La variance estimée de la distribution *a posteriori* du paramètre θ est alors comprise entre celle obtenue en regroupant les données des différentes études et celle calculée en supposant les études indépendantes. Plus les études sont homogènes, plus les estimations des

paramètres des différentes études bénéficient de l'information apportée par l'ensemble des études. Les moyennes des distributions *a posteriori* des θ_j tendent alors à se rapprocher de la moyenne de la distribution *a posteriori* de θ. Ce phénomène de rétrécissement *(shrinkage)* est plus important pour les petites études et pour celles dont les paramètres observés sont les plus éloignées de la moyenne générale.

Schémas adaptatifs

Les schémas d'études statiques classiques sont satisfaisants en termes d'inférence statistique mais peuvent présenter une limite à l'évaluation d'un nombre important de traitements potentiellement efficaces compte tenu de la nécessité de disposer du critère de jugement chez un grand nombre de patients avant l'analyse des résultats. Ce nombre de patients est déterminé avant le début de l'étude à partir des risques α et β, de l'effet à mettre en évidence θ et de certains paramètres de nuisance comme la variabilité du critère de jugement, la proportion de perdus de vue… qu'il est impossible de réviser en cours d'étude. Une mauvaise spécification de ces paramètres peut cependant entraîner un manque de puissance ou une durée d'essai trop longue. Des schémas adaptatifs ont été proposés, dans lesquels les données accumulées dans l'essai sont utilisées pour modifier certains aspects du plan d'expérience, sans en compromettre les résultats. Des schémas construits sur des méthodes fréquentistes existent, mais, de par la nature séquentielle de ces schémas, les approches bayésiennes montrent un certain nombre d'avantages dans ce cadre.

Réévaluation du nombre de sujets nécessaires en cours d'étude

L'approche bayésienne semble particulièrement adaptée à la question de l'information potentielle apportée par l'inclusion de *n* patients supplémentaires, en prenant en compte l'information disponible après inclusion de m patients dans l'étude. Il semble alors raisonnable d'estimer la puissance prédite, *e.g.* la probabilité prédictive de supériorité du bras test après inclusion des *n* patients supplémentaires, en basant les prédictions sur la distribution courante des résultats de l'étude [3]. Le calcul de cette probabilité prend en compte l'augmentation de l'incertitude sur θ découlant des prédictions.

Les essais séquentielles par groupe

Les analyses intermédiaires peuvent conduire à arrêter l'étude avant l'inclusion du nombre de sujets initialement prévus. Des plans expérimentaux ont été développés à partir de l'approche fréquentiste, autorisant des arrêts précoces pour efficacité, futilité ou toxicité [6-10]. Dans l'approche bayésienne, les résultats des analyses successives contribuent à la mise à jour des probabilités des hypothèses H_0 ou H_1, et les conséquences des observations à venir sur les résultats de fin d'étude sont pondérées par leurs probabilités bayésiennes prédictives. Cette approche fournit une réponse pragmatique aux différentes questions posées. Quelle estimation de la probabilité *a posteriori* de l'hypothèse d'intérêt emporterait la conviction ? Quels résultats obtenus durant l'étape à venir seraient compatibles avec cette estimation ? Quelle est la probabilité de ces résultats

compte tenu de l'information disponible ? L'approche bayésienne s'appuie ici sur le calcul de la somme des probabilités prédictives des résultats de la seconde étape qui conduiraient à une estimation de la probabilité de l'hypothèse d'intérêt supérieure à une valeur seuil prédéfinie, une fois l'étude terminée [5].

Les essais de phase I

En phase I, l'estimation de la dose maximale tolérée (DMT), *e.g.* correspondant à une probabilité prédéfinie de présenter une toxicité dose-limitante, s'appuie sur la mise en œuvre de schémas adaptatifs. En dehors des valeurs de DMT basses, le schéma d'escalade de dose non paramétrique classique 3+3 apparaît moins performant que la méthode de réévaluation séquentielle (*continual reassessment method*, CRM) [11] pour estimer la DMT [12]. Cette méthode repose sur une modélisation paramétrique de la relation dose-toxicité avec une attribution de dose et une estimation séquentielle du paramètre du modèle, jusqu'à l'inclusion d'un nombre de sujets fixe et préétabli ou jusqu'à ce qu'une règle d'arrêt soit vérifiée. La CRM a fait et fait encore aujourd'hui l'objet de nombreux travaux de recherche, en particulier pour l'estimation de la fonction dose-risque sous-jacente [13].

Les essais de phase II

En phase II, la recherche d'une dose optimale est rarement effectuée en cancérologie. L'efficacité thérapeutique est évaluée en phase II chez des patients recevant la DMT, considérée comme la dose la plus efficace. L'utilisation fréquente de polychimiothérapies a justifié le développement de nouveaux schémas d'étude. Huang *et al.* [14] ont ainsi proposé d'évaluer simultanément la sécurité et l'efficacité d'associations thérapeutiques dans un essai combiné de phase I-II. Ce schéma adaptif inclut une phase initiale d'escalade de dose, puis une randomisation adaptative des combinaisons de doses admissibles. La randomisation adaptative privilégie les niveaux de dose les plus efficaces à partir des valeurs des probabilités *a posteriori*. L'essai est arrêté si la probabilité *a posteriori* de toxicité, d'efficacité ou de futilité dépasse un seuil préalablement défini. Comparé à la mise en œuvre de deux essais distincts, le schéma combiné permet d'inclure moins de patients, est plus puissant et alloue davantage de patients à des doses plus efficaces. Le risque de négliger d'éventuelles toxicités tardives a cependant été soulevé récemment, et l'utilisation de ces méthodes doit être envisagée en fonction de situations spécifiques [15].

Les essais de phase II-III

Les essais combinés de phase II-III visent à éviter le cloisonnement usuel entre les phases II et III, potentiellement à l'origine de délais importants. Le schéma adaptatif proposé par Inoue *et al.* [16] intègre les phases II et III. Les patients sont randomisés entre les deux groupes de traitement dans un nombre restreint de centres en phase II. À chaque évaluation, la décision d'interrompre l'essai pour efficacité, de l'interrompre pour futilité ou de passer en phase III dépend des estimations des probabilités prédictives d'observer une meilleure survie dans le bras test une fois le suivi

réalisé. Le mélange de modèles paramétriques utilisé autorise des différences d'effets du traitement sur le taux de mortalité en fonction de l'obtention d'un contrôle local de la tumeur. Des simulations comparant le schéma proposé à une approche séquentielle groupée usuelle en phase III indiquent le gain de l'approche combinée en termes de durée d'étude et de nombre de sujets inclus.

À côté de ce schéma complètement bayésien, d'autres schémas « hybrides » pour des essais de phase II-III ont été proposés avec, par exemple, une analyse intermédiaire bayésienne mais une analyse finale « fréquentiste » afin de prouver le contrôle du risque d'erreur de première espèce qui est une condition nécessaire imposée par les autorités de régulation pour les essais pivot [17]. L'utilisation de méthodes bayésiennes a aussi été proposée dans un schéma de phase II-III pour des thérapies ciblées en cancérologie, avec sélection d'une sous-population à l'analyse intermédiaire [18]. Lors de l'analyse finale, les tests sont ajustés de façon à contrôler le risque d'erreur de première espèce de l'ensemble de l'essai.

Les méthodes de randomisation adaptatives

Les méthodes de randomisation adaptatives consistent à déséquilibrer la randomisation en donnant plus de poids au traitement se révélant le plus efficace. La probabilité d'allocation dans l'un des bras de l'étude évolue continuellement en fonction des distributions *a posteriori* des probabilités de succès des traitements candidats [19]. Une probabilité d'efficacité thérapeutique trop basse conduit à abandonner le traitement concerné. Cette méthode est applicable à la comparaison de plus de deux traitements [20] et permet la prise en compte de covariables [21].

Conclusion

Analysons, à titre d'illustration, les résultats de deux essais fictifs de phase II visant à estimer, dans une nouvelle indication, la réponse à une chimiothérapie d'efficacité connue dans des indications voisines. Les résultats obtenus en utilisant des approches fréquentiste et bayésienne sont comparés pour une proportion de répondeurs observée de 60 % chez respectivement 30 (essai n° 1) et 50 patients (essai n° 2).

Dans l'approche fréquentiste, la vraisemblance $p(y|\theta)$ est maximale pour des valeurs de θ égales aux proportions observées, soient respectivement 18/30 = 60 % *(figure 1a)* et 30/50 = 60 % *(figure 1b)*. L'intervalle de confiance de la proportion de répondeurs au niveau de confiance $1 - \alpha = 95\,\%$ est estimé à :

$$0{,}60 \pm 1{,}96\sqrt{\frac{0{,}60(1-0{,}60)}{30}}$$

soit [42,4 %-77,6 %] pour le premier essai, et

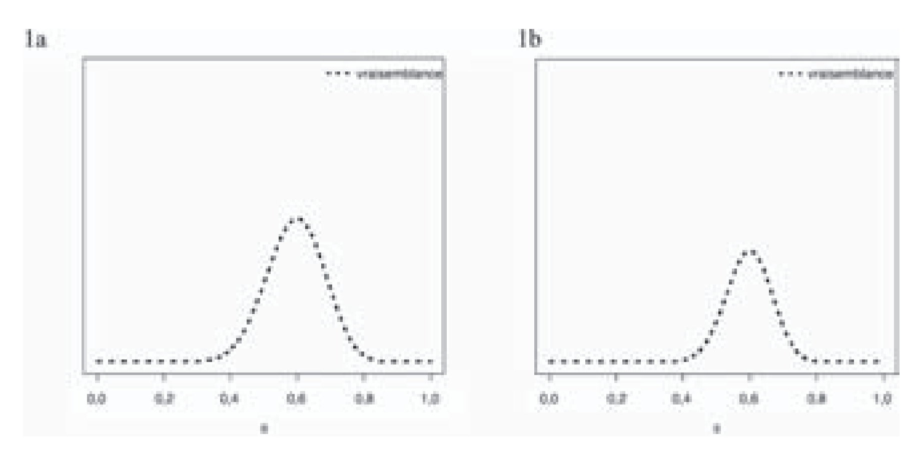


Figure 1. Vraisemblance $p(y|\theta)$ des essais n° 1 (figure 1a) et n° 2 (figure 1b).

$$0{,}60 \pm 1{,}96\sqrt{\frac{0{,}60(1-0{,}60)}{30}}$$

soit [46,4 %-73,6 %] pour le deuxième essai. Dans l'approche fréquentiste, l'analyse s'appuie sur l'information apportée par les seules données. Les données du deuxième essai apportent davantage d'information, l'intervalle de confiance est plus étroit.

Compte tenu des résultats obtenus dans des indications voisines, le niveau de réponse thérapeutique attendu *a priori* est estimé à 30 %. Le niveau d'incertitude sur la distribution $p(\theta)$ est résumé en considérant que celui-ci correspond à l'information qu'aurait apportée une étude réalisée chez 30 sujets. Cette information définit la loi de probabilité *a priori* $p(\theta)$, loi bêta dont la densité est représentée sur les *figures 2a et 2b*. Pour le premier essai, l'estimation de la distribution de probabilité *a posteriori* $p(\theta|y)$ présente un mode à 44,8 %, et l'intervalle de crédibilité à 95 % de θ s'étend de 32,7 % à 57,6 %. Ainsi, la prise en compte de l'information *a priori* sur la proportion attendue de répondeurs a conduit à tempérer les « trop bons » résultats issus de l'analyse des données. Pour le deuxième essai, la distribution de probabilité *a posteriori* $p(\theta|y)$ présente un mode à 48,7 %, l'intervalle de crédibilité à 95 % de θ s'étend de 37,9 % à 59,7 %. L'information apportée par les données est ici plus importante, ces dernières contribuent davantage à l'estimation de la distribution de $p(\theta|y)$. Cette plus grande précision de l'information apportée par les données conduit à une estimation de la distribution *a posteriori* plus proche de la proportion observée et à un intervalle de crédibilité plus étroit.

Les investigateurs peuvent donner davantage de poids à leur *a priori*, en fonction du niveau d'incertitude qu'ils ont sur la valeur de θ. Pour un même niveau de réponse thérapeutique attendu de 30 %, l'information qu'aurait apportée une étude réalisée chez 100 sujets correspond à davantage de précision. Cette information *a priori* définit la loi de $p(\theta)$, loi bêta dont la densité est représentée sur les *figures 3a et 3b*. Pour le premier essai, la distribution de probabilité *a posteriori* $p(\theta|y)$ présente un mode à 36,7 %, l'intervalle de crédibilité à 95 % de θ s'étend de 28,8 % à 45,4 % *(figure 3a)*. Le poids plus important donné à l'information *a priori* conduit à rapprocher

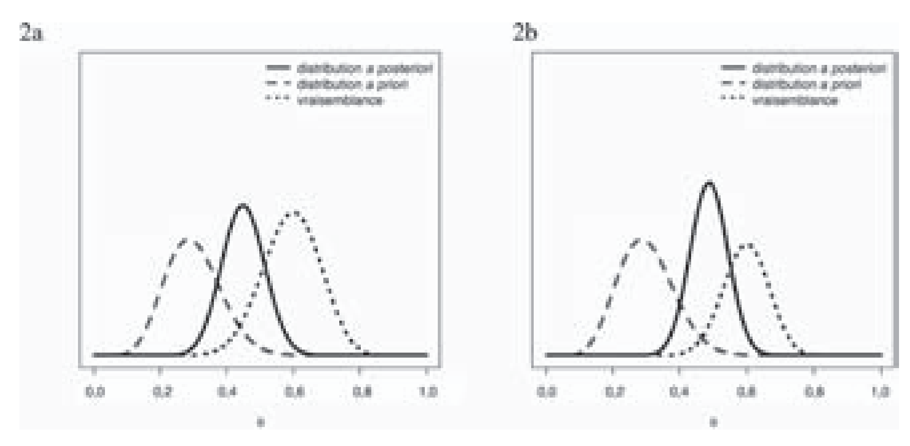


Figure 2. Distribution *a priori* $p(\theta)$, vraisemblance $p(y|)$ et distribution *a posteriori* $p(\theta|y)$ du paramètre des essais n° 1 (figure 2a) et n° 2 (figure 2b).

l'estimation de la distribution *a posteriori* de θ de l'information historique. Pour le deuxième essai, la distribution de probabilité *a posteriori* $p(\theta|y)$ présente un mode à 39,9 %, l'intervalle de crédibilité à 95 % de θ s'étend de 32,3 % à 48,0 % *(figure 3b)*. La précision de l'information *a priori* tempère ici aussi l'optimisme issu de l'analyse des données de l'étude courante.

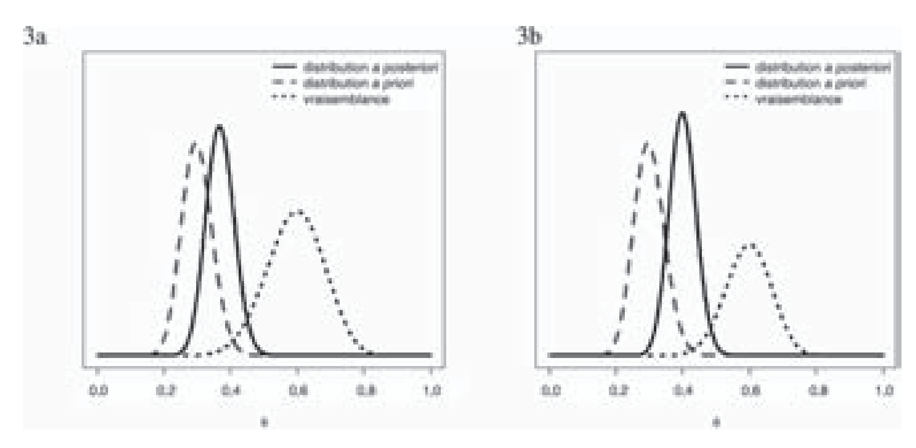


Figure 3. Distribution *a priori* $p(\theta)$, vraisemblance $p(y|\theta)$ et distribution *a posteriori* $p(\theta|y)$ du paramètre des essais n° 1 (figure 3a) et n° 2 (figure 3b).

Références

1. Adamina M, Tomlinson G, Guller U. Bayesian statistics in oncology. *Cancer* 2009 ; 115 : 5371-81.

2. Hoenig JM, Heisey DM. The Abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician* 2001 ; 55 (1) : 19-24.

3. Spiegelhalter DJ, Abrams KR, Myles JP. *Bayesian approaches to clinical trials and health-care evaluation*. New York: John Wiley & Sons, 2004, 406 pages.

4. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian data analysis*. Second edition. London : Chapman & Hall, 2004, 668 pages.

5. Berry DA. Statistical innovations in cancer research. In : Holland J, Frei T, *et al.* (eds). *Cancer Medicine* London : BC Decker, 2003 (e.6. Ch 33), pp. 465-78.

6. Gehan EA. The determination of the number of patients in a follow-up trial of a new chemotherapeutic agent. *J Chron Dis* 1961 ; 13 : 346-53.

7. Fleming TR. One sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982 ; 38 : 143-51.

8. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989 ; 19 : 1-10.

9. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979 ; 35 : 549-55.

10. Whitehead J. *The design and analysis of sequential clinical trials*. Revised second edition. New York : John Wiley & Sons, 1992, 314 pages.

11. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* 1990 ; 46 : 33-48.

12. Iasonos A, Wilton AS, Riedel ER, Seshan VE, Spriggs DR. A comprehensive comparison of the continual reassessment method to the standard 3+3 dose escalation scheme in phase I dose-finding studies. *Clin Trials* 2008 ; 5 : 465-77.

13. Daimon T, Zohar S, O'Quigley J. Posterior maximization and averaging for Bayesian working model choice in the continual reassessment method. *Stat Med* 2011 ; 30 : 1563-73.

14. Huang X, Biswas S, Oki Y, Issa JP, Berry DA. A parallel phase I/II clinical trial design for combination therapies. *Biometrics* 2007 ; 63 : 429-36.

15. Table ronde N° 2 – Les méthodes adaptatives : quand et comment les utiliser dans les essais cliniques ? Les Ateliers de Giens. Rencontres nationales de pharmacologie clinique, 10-11 janvier 2011. http://www.ateliersdegiens.org/ *Thérapie* 2011 ; 66 : 309-17.

16. Inoue LYT, Thall PF, Berry DA. Seamlessly expanding a randomized phase II trial to phase III. *Biometrics* 2002 ; 58 : 823-31.

17. Schmidli H, Bretz F, Racine Poon A. Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Stat Med* 2007 ; 26 : 4925-38.

18. Brannath W, Zuber E, Branson M, *et al.* Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Stat Med* 2009 ; 28 : 1445-63.

19. Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. *Eur J Cancer* 2007 ; 43 : 859-66.

20. Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov* 2006 ; 5 : 27-36.

21. Thall PF, Wathen JK. Covariate-adjusted adaptive randomization in a sarcoma trial with multi-stage treatments. *Stat Med* 2005 ; 24 : 1947-64.

Partie II
Critères de jugement

Critères de réponse

R. Porcher, A. Kramar

Le but des essais de phase II est d'identifier les traitements prometteurs pour un type de tumeur donné. Le critère de jugement est classiquement un critère binaire (succès/échec), le succès (ou réponse au traitement) étant défini par la disparition ou une réduction importante de la taille de la tumeur après une certaine durée de traitement. Il est aujourd'hui acquis que, pour un certain nombre de tumeurs solides, un traitement qui entraîne une diminution de la masse tumorale a une probabilité raisonnable d'entraîner un allongement de la survie, comme par exemple dans le cancer colorectal [1], mais cela n'a pas été démontré pour tous les types de tumeur.

En tant que critère de jugement principal dans la plupart des essais de phase II en cancérologie, une définition précise de la réponse était nécessaire, notamment pour permettre une évaluation homogène entre les études. Des critères standardisés ont donc été développés, avec une catégorisation de la réponse en quatre classes [2, 3]. Après des décennies d'utilisation, ces critères ont récemment fait l'objet de critiques, notamment parce qu'ils ne tenaient pas compte de techniques innovantes ou parce qu'ils n'étaient pas adaptés à l'évaluation de certaines classes de médicament.

Critères de réponse catégoriels

Critères OMS

Après les premiers efforts pour définir la réponse d'une tumeur à un agent anticancéreux dans les années 1960 à 70, les critères de l'Organisation mondiale de la santé (OMS) ont été développés pour permettre une évaluation et surtout un rapport des résultats de manière standardisée [2]. Ces critères utilisent la variation relative de taille de la tumeur entre une mesure avant traitement et une mesure après l'exposition au traitement, par exemple après 2 ou 4 cycles de chimiothérapie. Quatre catégories de réponse sont définies : disparition complète de la tumeur (réponse complète, RC), diminution d'au moins 50 % de la taille de la tumeur (réponse partielle, RP), augmentation d'au moins 25 % de la taille de la tumeur ou apparition de nouvelles lésions (maladie progressive, MP), variation de la taille de la tumeur ne permettant de la classer ni en réponse ni en progression (maladie stable, MS).

La taille (ou volume) de la tumeur est estimée par le produit du plus grand diamètre et le diamètre perpendiculaire si la tumeur est considérée mesurable dans deux dimensions. Dans le cas de

plusieurs lésions présentes dans le même organe cible, la taille totale reflétant la masse tumorale est obtenue par la somme des produits des diamètres de toutes ces lésions. Si la tumeur n'est pas mesurable dans deux dimensions, comme par exemple une adénopathie médiastinale, seule la plus grande dimension est considérée. Enfin, pour des tumeurs non mesurables à l'aide d'une règle ou d'un compas, comme des métastases osseuses ou des masses pelviennes ou abdominales, la variation relative doit être simplement estimée.

En pratique, dans la mesure où avec des chimiothérapies cytotoxiques une stabilisation de la maladie est souvent temporaire, une maladie stable était plutôt considérée comme signifiant un échec, et une réponse objective (RO) dans un essai de phase II définie comme l'obtention d'une réponse partielle ou complète.

Critères RECIST

Si les critères OMS ont permis une homogénéité de présentation des résultats, plusieurs questions se sont posées avec leur application en pratique, notamment à cause d'erreurs ou de difficultés de mesure. En particulier la façon d'intégrer dans l'évaluation de la réponse les variations de taille des lésions mesurables et non mesurables, ainsi que la taille minimale et le nombre de lésions évaluées variait considérablement selon les études. Par ailleurs, les critères OMS ont parfois été adaptés à des situations particulières ou pour tenir compte de nouvelles technologies comme le scanner ou l'imagerie par résonance magnétique (IRM). Enfin, selon les études, la maladie pouvait être considérée comme progressive en fonction de l'augmentation de taille d'une seule lésion ou de la somme de la taille des lésions. Afin de pallier ces problèmes, des critères dits RECIST *(Response Evaluation Criteria in Solid Tumors)* ont été publiés en 2000 [3] et mis à jour tout récemment (RECIST 1.1) [4].

Le guide d'évaluation RECIST se divise en deux parties : d'une part les critères de réponse proprement dits, qui reprennent une évaluation catégorielle de la réponse sur le même mode que les critères OMS (RC/RP/MS/MP) et, d'autre part, des recommandations pour l'évaluation et la mesure des lésions, notamment en fonction des techniques utilisées et de leur taille.

Par rapport aux critères OMS, les critères RECIST différencient des lésions, dites lésions cibles au nombre maximal de 5 (10 dans la version originale), qui sont mesurées et des lésions non cibles, suivies mais non mesurées. Les lésions cibles sont choisies en fonction de leur taille (lésions de plus grands diamètres) et de leur localisation représentative des organes touchés (pas plus de 2 par organe). La mesure de la taille des lésions devient unidimensionnelle et non plus bidimensionnelle : seul le plus long diamètre des lésions est mesuré, et la taille totale dont la variation est analysée correspond à la somme des plus grands diamètres des lésions cibles. Le choix s'est porté sur une mesure unidimensionnelle pour des raisons de simplicité d'une part, mais aussi parce que ce critère a été considéré comme ayant une relation plus proche de la linéarité avec le logarithme du nombre de cellules tumorales (et donc la masse tumorale) que le produit des deux plus grands diamètres, au moins pour des tumeurs sphériques [5]. Des critères de mesurabilité sont aussi introduits : la taille minimale des lésions mesurables est de 10 mm dans RECIST 1.1,

que ce soit une mesure obtenue par scanner ou clinique (à l'aide d'un compas). Depuis RECIST 1.1, les ganglions de plus de 10 mm sont aussi pris en compte [4], alors qu'ils n'étaient pas mentionnés dans la version originale.

Les seuils pour définir une réponse partielle ou la progression diffèrent par rapport aux critères OMS : une diminution d'au moins 30 % de la somme des plus grands diamètres des **lésions cibles** par rapport à l'inclusion définit une RP, et une augmentation d'au moins 20 % de la somme des plus grands diamètres par rapport à la valeur la plus basse recueillie depuis le début de l'étude classe les patients en MP, si cette variation est supérieure à 5 mm en valeur absolue. Si les seuils sont différents, il faut noter que, pour des tumeurs sphériques, le seuil de 30 % retenu pour la RP est à peu près équivalent à la diminution de surface de 50 % des critères OMS. Néanmoins, il a été noté que plus de réponses sont classées partielles avec RECIST qu'avec OMS. Pour les **lésions non cibles**, une RC signifie la disparition de toutes les lésions et la normalisation des marqueurs tumoraux (dans les situations où ces marqueurs ont été validés), les ganglions étant considérés comme normaux en dessous de 10 mm ; le classement en MP nécessite une aggravation substantielle de la lésion, appelée progression non équivoque, afin d'éviter une surestimation du taux de progression en l'absence de mesure objective, ou l'apparition de nouvelles lésions. Enfin, entre ces deux catégories, aucune RP ne peut être définie, mais simplement un classement en ni RC/ni MP. L'évaluation globale du patient en fonction des réponses des lésions cibles et des lésions non cibles est détaillée dans l'article original [4].

Afin d'éviter de surestimer le taux de réponse, les critères OMS et RECIST prévoyaient que les variations de taille des lésions devaient être confirmées par une deuxième mesure réalisée au moins à 1 mois d'intervalle [2, 3]. Les critères RECIST 1.1 ne conservent cette mesure de confirmation que pour les essais où la réponse est le critère de jugement principal [4].

Du point de vue pratique, de nombreuses études ont comparé les critères OMS et RECIST, et aucune différence majeure sur les taux de réponse n'a été trouvée, alors qu'avec les critères RECIST les taux de progression sont légèrement plus faibles, et les délais de progression légèrement plus longs [6]. Une étude plus théorique tenant compte de la géométrie des tumeurs a cependant montré que des discordances plus substantielles pouvaient survenir [7]. Un exemple de calcul de la réponse à partir des critères RECIST est présenté dans le *tableau I*.

Si de nouvelles lésions étaient apparues, le patient de cet exemple aurait été considéré en progression. Si la diminution de taille des lésions avait été inférieure à 30 % (par exemple, si la taille de la lésion 2 du foie reste constante), la réponse aurait été considérée comme stable.

La ***durée de la réponse*** est aussi un critère d'évaluation important pour étudier l'effet d'un traitement. L'obtention d'une RC de courte durée, suivie d'une récidive de la tumeur peut être considérée comme un échec du traitement, alors qu'une RP, voire une MS sur une longue durée peut indiquer un succès. Par le passé, de nombreuses définitions de la durée de la réponse ont été utilisées, en traitant parfois les RC et les RP de façon différente. Les critères RECIST définissent la durée de réponse globale comme la durée entre la date où le premier critère de réponse est rempli (que ce soit RP ou RC) et la date où la récidive ou la progression est documentée

Tableau I. Exemple de calcul de la réponse selon les critères RECIST pour un patient présentant des lésions hépatiques d'une tumeur primitive et des métastases cérébrales.

Organe		Lésion cible	Taille avant traitement (mm)	Taille après traitement (mm)	Variation relative (%)
Foie	Lésion 1	x	30	20	
	Lésion 2	x	20	13	
	Lésion 3		12	10	
Cerveau	Lésion 1	x	10	5	
	Lésion 2		5	5	
Nouvelles lésions				Non	
Somme de la taille des lésions cibles			60	38	- 37 %
Réponse des lésions cibles					RP
Réponse des autres lésions					Ni RC/ni MP
Réponse globale					RP

objectivement. La durée de réponse complète, quant à elle, est définie à partir de l'obtention d'une RC. Enfin, la durée de maladie stable est définie à partir de la date de début de traitement (ou de randomisation si l'essai est randomisé).

Critères adaptés à certaines tumeurs

Malgré les progrès apportés par les critères RECIST dans la définition d'une tumeur mesurable et la façon de mesurer les lésions, ces critères ne sont pas adaptés à tous les types de tumeurs et il reste des domaines où des critères différents sont utilisés [6, 8]. Dans les tumeurs cérébrales, par exemple, beaucoup préfèrent des mesures bi- ou tridimensionnelles [9], en dépit d'une bonne concordance entre ces différentes évaluations dans la détection des réponses [6].

D'autres travaux ont en revanche montré que les critères RECIST n'étaient pas adaptés à l'évaluation de la réponse au traitement dans les tumeurs stromales gastro-intestinales (GIST), pour lesquelles la taille de la tumeur peut augmenter sous l'effet d'une réponse métabolique à certains agents ou rester stable malgré une résistance au traitement [10]. Des critères spécifiques combinant des variations de taille et de densité mesurée au scanner ont alors été proposés [11]. De même, la capacité à mesurer des tumeurs métastatiques de la prostate est un problème, et

l'évaluation de la réponse dans ce type de cancers nécessite le développement de nouveaux outils intégrant des marqueurs comme les PSA, les critères RECIST si la tumeur est mesurable, une évaluation osseuse et histologique, ainsi que la qualité de vie [12].

Enfin, des critères (dit critères de CHESON) ont été aussi présentés pour des tumeurs non solides comme le lymphome, qui reprennent la classification RC/RP/MS/MP avec des définitions différentes, intégrant des mesures en imagerie par tomographie par émission de positons (TEP) et scanner, la palpation de la rate et du foie, ainsi que le résultat de biopsies médullaires [13].

Présentation des résultats

Le rapport de l'évaluation de la réponse est généralement descriptif, avec la présentation du nombre de patients classé dans les catégories RC, RP, MS, MP au temps protocolaire d'évaluation. Parfois, la meilleure réponse obtenue est aussi rapportée. Les nombres de patients inéligibles et non évaluables pour la réponse (par exemple, parce qu'ils sont décédés ou sortis prématurément de l'étude) doivent aussi être donnés, ainsi que les motifs de ces classements. Le taux de réponse objective est présenté habituellement avec un intervalle de confiance à 95 % bilatéral. Ce pourcentage dépend du dénominateur choisi, et il est important que le mode de calcul soit fixé dans le protocole et rapporté. Les critères RECIST recommandent d'inclure tous les patients éligibles dans le dénominateur. Une stratégie plus proche du principe de l'intention de traiter consiste à considérer tous les patients inclus. Pour ces calculs, les patients non évaluables sont considérés comme non répondeurs. Un exemple de l'influence du mode de calcul est donné dans le *tableau II* à partir des résultats d'un essai fictif ayant inclus 34 patients, dont 2 inéligibles et 4 non évaluables.

Tableau II. Taux de réponse obtenu en fonction de la population analysée.

	Patients inclus (n = 34)	**Patients éligibles (n = 32)**	**Patients éligibles et évaluables (n = 28)**
Réponse complète	1	1	1
Réponse partielle	7	7	7
Maladie stable	11	11	11
Maladie progressive	9	9	9
Non évaluable	4	4	–
Taux de réponse [IC 95 %]	24 % [11 % ; 41 %]	25 % [11 % ; 43 %]	29 % [13 % ; 49 %]

Autres critères de réponse

Limitations et critiques des critères catégoriels de régression tumorale

De nombreuses critiques se sont élevées récemment contre les critères de réponse présentés en catégories [8, 14, 15].

En premier lieu, la catégorisation d'un critère continu comme le pourcentage de régression tumorale entraîne une perte d'information qui peut être importante [16]. Par exemple, les réponses considérées comme stables selon RECIST comportent à la fois des tumeurs qui augmentent de 15 % ou qui diminuent de 25 %. D'un point de vue statistique, cette catégorisation entraîne aussi une perte de puissance dont la seule justification est la facilité de regroupement en catégories [17].

Par ailleurs, les critères de régression de la masse tumorale ont été développés dans le cadre de l'évaluation d'agents cytotoxiques. Avec le développement de molécules non cytotoxiques, il est apparu qu'un médicament pouvait être actif et prolonger la survie ou la survie sans progression sans provoquer de régression importante de la taille de la tumeur à court terme [14, 18, 19]. Parfois, un ralentissement de la croissance tumorale peut aussi déjà indiquer une réponse au traitement [19].

D'autres moyens de décrire la réponse au traitement ont donc été explorés pour pallier ces difficultés. Il faut cependant aussi noter dans un certain nombre de ces critiques une confusion entre le choix d'un critère de jugement pour un essai de phase II et la manière de mesurer une tumeur (RECIST) [15]. En effet, rien n'empêche, par exemple, de considérer que l'obtention d'une stabilisation de la maladie correspond à une « réponse » pour l'évaluation d'un traitement particulier dans une maladie donnée. De plus, ces critiques sont aussi liées à un constat d'inefficacité relative des essais de phase II dans l'identification des molécules prometteuses pour des essais de phase III [14]. Au-delà d'une critique de la mesure de la réponse d'une tumeur à un traitement, il s'agit en fait plus d'une critique du schéma traditionnel des essais de phase II non comparatifs. À ce titre, des essais de phase II randomisés avec des critères comme la survie sans progression semblent une piste envisagée aujourd'hui par de nombreux auteurs [20-23]. Il faut alors noter que si la survie sans progression est utilisée comme critère de réponse, la fréquence des examens permettant d'évaluer la progression est un paramètre important : des examens trop espacés tendront à surestimer la durée de la survie sans progression (*cf.* chapitre II.3 « La survie comme critère de jugement », page 85). Dans le cadre de l'évaluation d'un essai randomisé en aveugle cependant, le choix du critère d'évaluation se révèle moins crucial que pour un essai non comparatif.

Enfin, des problèmes liés à l'évaluation pratique de ces critères de réponse subsistent. D'une part, il faut prévoir comment considérer les patients qui n'ont pas de mesure de confirmation ou pour lesquels l'évaluation de confirmation a lieu avant ou bien au-delà d'un mois après la première évaluation. Par ailleurs, la prise en compte de nouvelles lésions est l'objet de controverses. Selon les critères OMS ou RECIST, elles classent le patient en MP, quels que soient leur nombre et leur taille. Ce critère peut être jugé comme trop strict si de nouvelles petites lésions apparaissent alors

que les autres lésions répondent clairement [18] ou on peut, au contraire, considérer que de nouvelles lésions, mêmes petites, sont très importantes. Certains auteurs ont proposé d'ajouter la taille des nouvelles lésions à la taille totale des lésions, lorsqu'elles étaient mesurables [24, 25].

Évolution de la masse tumorale

À côté de l'analyse d'une réponse tumorale présentée en catégories, une représentation graphique des variations de taille des lésions observées dans l'étude par des graphes en cascade (*waterfall plots*) [26] est aujourd'hui couramment utilisée. Un exemple est présenté dans la *figure 1* à partir de données fictives.

Ces graphes permettent de visualiser de façon détaillée les résultats obtenus en termes de régression ou d'augmentation de la taille des tumeurs, et la survenue de nouvelles lésions peut être prise en compte en différenciant sur le graphique les patients considérés comme en réponse partielle ou avec une maladie stable des patients en progression. L'inconvénient de ces graphes est, d'une part, leur nature descriptive plutôt qu'analytique et, d'autre part, la difficulté de représenter les patients non évaluables ou avec une évaluation manquante au cours du suivi.

Pour éviter la perte d'information liée à la catégorisation de la réponse, plusieurs travaux ont aussi défendu une analyse quantitative de la variation de la taille tumorale [17, 24]. Ce type d'analyse présente en outre l'avantage de pouvoir tenir compte de covariables associées à la régression tumorale, voire de modéliser les mesures répétées au cours du suivi par une approche longitudinale. En revanche, cette approche est mieux adaptée à des essais randomisés qu'à des essais monobras, dans la mesure où la comparaison de la variation de taille de la tumeur est importante.

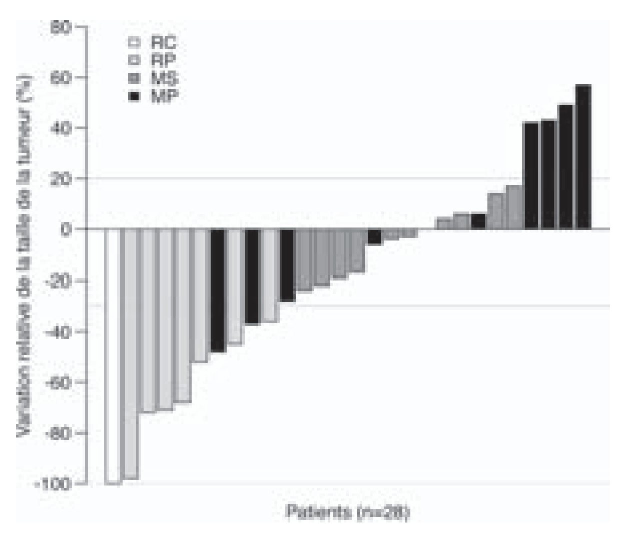


Figure 1. Exemple de graphe en cascade *(waterfall plot)* sur les données fictives du tableau II. Des patients dont la tumeur régresse de plus de 30 % ou n'augmente pas de plus de 20 % peuvent être considérés comme en progression si de nouvelles lésions apparaissent.

Marqueurs tumoraux

Dans certains cas, la réponse à un traitement peut aussi être définie à partir de mesures de marqueurs tumoraux validés, afin d'éviter une irradiation répétée des patients par des scanners ou lorsque la réponse clinique est difficile à évaluer. C'est par exemple le cas pour le cancer de la prostate, où l'évaluation clinique et scanographique est compliquée [12] et où un marqueur, l'antigène prostate spécifique (PSA) est un bon indicateur de la masse tumorale.

Plusieurs auteurs ont donc cherché à définir des critères de réponse utilisant des mesures répétées de marqueurs tumoraux, à partir de la corrélation entre les variations de ces marqueurs et des critères de réponse clinique. C'est le cas par exemple de critères de réponse basés sur la mesure du CA125 dans le cancer de l'ovaire recevant une première ligne de traitement [27]. Ces critères ressemblent néanmoins aux critères de réponse clinique (OMS, RECIST) dans la mesure où les catégories sont définies à partir de seuils de la variation du marqueur depuis l'inclusion.

Une évaluation de la réponse à partir de marqueurs tumoraux pose aussi des problèmes méthodologiques, comme l'évaluation clinique peut poser des problèmes liés à la catégorisation des réponses et à la fréquence des évaluations. En particulier, il est important de déterminer la fréquence des mesures du marqueur et la façon de traiter les mesures qui oscillent d'un prélèvement à l'autre. En plus de la question de la variabilité biologique et de l'erreur de mesure, il faut aussi veiller à ce que la mesure de référence ne soit pas trop éloignée du début du traitement, afin de ne pas classer par erreur une tumeur comme progressive, alors que le marqueur diminue sous traitement, après avoir augmenté entre la valeur de référence et l'instauration du traitement.

Enfin, très récemment, une stratégie innovante d'évaluation de la réponse a été proposée à partir de la modélisation mathématique de l'évolution d'un marqueur, en l'occurrence les PSA dans le cancer de la prostate [28].

Critères de réponse adaptés à l'évaluation d'agents non cytotoxiques

Une série de rencontres scientifiques sur le thème des agents immunothérapeutiques dans le cancer en 2004-2005 a permis de mettre en évidence les points suivants [18] :

- l'apparition d'un effet antitumoral mesurable peut prendre plus de temps qu'avec des agents cytotoxiques ;
- la réponse peut survenir après une phase de croissance tumorale entraînant un classement en MP selon les critères traditionnels ;
- l'apparition de nouvelles petites lésions est possible alors que les lésions plus grosses répondent, ce qui entraînerait un classement de MP avec les critères RECIST ;
- une MS durable peut représenter une activité antitumorale significative.

Des critères de réponse *ad hoc* pour l'évaluation de ce type de traitement ont alors pu être proposés, comme les critères irRC *(immune-related Response Criteria)*, développés à partir des critères OMS [25].

Place de l'imagerie

La place des méthodes récentes d'imagerie volumétrique ou fonctionnelle dans l'évaluation de la réponse, et en particulier celle de l'IRM et de la TEP au 18FDG, s'est posée ces dernières années, que ce soit en plus ou à la place de l'évaluation anatomique des critères RECIST [29, 30]. Dans sa révision 1.1, le groupe de travail RECIST a cependant noté que, malgré leur côté prometteur, il manquait encore à ces techniques une validation clinique rigoureuse et appropriée pour les intégrer dans les critères de réponse [4].

À retenir

Les critères de réponse dans les essais de phase II en cancérologie sont le plus souvent des critères catégoriels définis à partir de la variation relative de la taille de la tumeur entre une mesure avant traitement et une mesure après traitement (critère OMS et critère RECIST). Quatre catégories de réponse sont considérées : réponse complète (RC), réponse partielle (RP), maladie stable (MS) et maladie progressive (MP). Ces critères ont fait l'objet de critiques récentes, en particulier parce qu'ils n'étaient pas toujours adaptés à l'évaluation de thérapies innovantes et qu'ils ne tenaient pas compte des progrès récents en imagerie ou en biologie. Enfin, des critères comme la survie sans progression commencent à être recommandés, dans le cadre d'une remise en cause plus profonde des schémas des essais de phase II monobras au profit d'essais de phase II randomisés (*cf.* chapitre II.3, page 85).

Références

1. Buyse M, Thirion P, Carlson RW, *et al.* Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: A meta-analysis. Meta-Analysis Group in Cancer. *Lancet* 2000 ; 356 : 373-8.

2. Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting results of cancer treatment. *Cancer* 1981 ; 47 : 207-14.

3. Therasse P, Arbuck SG, Eisenhauer EA, *et al.* New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst* 2000 ; 92 : 205-16.

4. Eisenhauer EA, Therasse P, Bogaerts J, *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* 2009 ; 45 : 228-47.

5. James K, Eisenhauer E, Christian M, *et al.* Measuring response in solid tumors : Unidimensional *versus* bidimensional measurement. *J Natl Cancer Inst* 1999 ; 91 : 523-8.

6. Therasse P, Eisenhauer EA, Verweij J. RECIST revisited: A review of validation studies on tumour assessment. *Eur J Cancer* 2006 ; 42 : 1031-9.

7. Mazumdar M, Smith A, Schwartz LH. A statistical simulation study finds discordance between WHO criteria and RECIST guideline. *J Clin Epidemiol* 2004 ; 57 : 358-65.

8. McHugh K, Kao S. Response evaluation criteria in solid tumours (RECIST): Problems and need for modifications in paediatric oncology? *Br J Radiol* 2003 ; 76 : 433-6.

9. Galanis E, Buckner JC, Maurer MJ, *et al*. Validation of neuroradiologic response assessment in gliomas: Measurement by RECIST, two-dimensional, computer-assisted tumor area, and computer-assisted tumor volume methods. *Neuro Oncol* 2006 ; 8 : 156-65.

10. Benjamin RS, Choi H, Macapinlac HA, *et al*. We should desist using RECIST, at least in GIST. *J Clin Oncol* 2007 ; 25 : 1760-4.

11. Choi H, Charnsangavej C, Faria SC, *et al*. Correlation of computed tomography and positron emission tomography in patients with metastatic gastrointestinal stromal tumor treated at a single institution with imatinib mesylate: Proposal of new computed tomography response criteria. *J Clin Oncol* 2007 ; 25 : 1753-9.

12. Scher HI, Morris MJ, Kelly WK, *et al*. Prostate cancer clinical trial end points: “RECIST”ing a step backwards. *Clin Cancer Res* 2005 ; 11 : 5223-32.

13. Cheson BD, Pfistner B, Juweid ME, *et al*. Revised response criteria for malignant lymphoma. *J Clin Oncol* 2007 ; 25 : 579-86.

14. Ratain MJ, Eckhardt SG. Phase II studies of modern drugs directed against new targets: If you are fazed, too, then resist RECIST. *J Clin Oncol* 2004 ; 22 : 4442-5.

15. Twombly R. Criticism of tumor response criteria raises trial design questions. *J Natl Cancer Inst* 2006 ; 98 : 232-4.

16. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* 2006 ; 25 : 127-41.

17. Lavin PT. An alternative model for the evaluation of antitumor activity. *Cancer Clin Trials* 1981 ; 4 : 451-7.

18. Hoos A, Parmiani G, Hege K, *et al*. A clinical development paradigm for cancer vaccines and related biologics. *J Immunother* 2007 ; 30 : 1-15.

19. Takimoto CH. Commentary : Tumor growth, patient survival, and the search for the optimal phase II efficacy endpoint. *Oncologist* 2008 ; 13 : 1043-5.

20. Rubinstein LV, Korn EL, Freidlin B, *et al*. Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol* 2005 ; 23 : 7199-206.

21. Parmar MK, Barthel FM, Sydes M, *et al*. Speeding up the evaluation of new agents in cancer. *J Natl Cancer Inst* 2008 ; 100 : 1204-14.

22. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: The importance of randomisation. *Eur J Cancer* 2009 ; 45 : 275-80.

23. Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 2009 ; 27 : 2874-80.

24. Karrison TG, Maitland ML, Stadler WM, Ratain MJ. Design of phase II cancer trials using a continuous endpoint of change in tumor size: Application to a study of sorafenib and erlotinib in non small-cell lung cancer. *J Natl Cancer Inst* 2007 ; 99 : 1455-61.

25. Wolchok JD, Hoos A, O’Day S, *et al*. Guidelines for the evaluation of immune therapy activity in solid tumors : Immune-related response criteria. *Clin Cancer Res* 2009 ; 15 : 7412-20.

26. Ratain MJ, Eisen T, Stadler WM, *et al*. Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. *J Clin Oncol* 2006 ; 24 : 2505-12.

27. Rustin GJ, Nelstrop AE, McClean P, *et al*. Defining response of ovarian carcinoma to initial chemotherapy according to serum CA 125. *J Clin Oncol* 1996 ; 14 : 1545-51.

28. Stein WD, Figg WD, Dahut W, *et al*. Tumor growth rates derived from data for patients in a clinical trial correlate strongly with patient survival: A novel strategy for evaluation of clinical trial data. *Oncologist* 2008 ; 13 : 1046-54.

29. Sargent DJ, Rubinstein L, Schwartz L, *et al*. Validation of novel imaging methodologies for use as cancer clinical trial end-points. *Eur J Cancer* 2009 ; 45 : 290-9.

30. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: Evolving considerations for PET response criteria in solid tumors. *J Nucl Med* 2009 ; 50 (Suppl 1) : 122S-50S.

Critères de tolérance

S. Gourgou-Bourgade, A. Kramar

La tolérance est un critère majeur recueilli de façon systématique dans les essais cliniques afin d'estimer le bénéfice/risque du traitement à l'étude et ce, tout au long du développement du médicament depuis la phase I jusqu'à la phase IV de surveillance à long terme. Le terme « médicament » est utilisé de manière générique, car il peut s'agir d'autres types d'interventions thérapeutiques.

Au cours des essais de phase I, souvent appelés essais de tolérance, la toxicité est le critère principal permettant de déterminer la dose recommandée pour la phase II. L'information recueillie est très précise, surtout quand il s'agit d'une première administration chez l'homme (*cf.* chapitre V.1 « Planification d'un essai de phase I », page 269).

Dans les essais de phase II, le critère principal est orienté pour détecter un signal d'efficacité, et la tolérance est également recueillie pour pouvoir mettre en balance les effets bénéfiques (*cf.* chapitre V.2 « Mise en œuvre d'un essai clinique de phase II », page 281). Certains plans expérimentaux permettent même d'utiliser ces deux critères simultanément pour décider de continuer ou non en phase III.

Dans les essais de phase III, étape qui cherche à démontrer une efficacité à plus long terme, la tolérance est recueillie de manière plus ciblée en vue de pondérer l'estimation de l'efficacité afin de juger globalement de la valeur thérapeutique du traitement évalué (*cf.* chapitre V.3 « Mise en œuvre d'un essai clinique de phase III », page 301).

Le recueil des données en phase IV et dans les études de pharmaco-épidémiologie permet d'évaluer le risque, le bénéfice et l'usage des médicaments ou d'une association médicamenteuse sur une population plus large une fois le médicament mis sur le marché.

L'objectif de ce chapitre est de définir les moyens mis en œuvre pour évaluer la tolérance des thérapies à l'étude afin de préserver la sécurité des individus à tout moment dans le développement des stratégies thérapeutiques.

Définitions

Afin d'évaluer la tolérance d'un médicament, il faut commencer par recueillir l'information qui a été préalablement définie dans le protocole. Les différentes toxicités d'ordre clinique et/ou biologique, plus communément appelées effets indésirables ou événements indésirables, peuvent être classées en plusieurs catégories selon leur gravité.

Effet indésirable

Un effet indésirable est une manifestation nocive, non désirée, survenant chez un patient traité ou ayant été traité par un médicament et qui est attribuée à ce dernier.

Événement indésirable

Un événement indésirable est un événement nocif et non recherché survenant chez un sujet exposé ou non à un traitement ou un facteur de risque donné. Un événement indésirable, selon l'art. R1123-39 du Code de la Santé publique, est défini comme un événement fâcheux d'ordre clinique et/ou biologique chez un individu participant à une étude clinique, à qui on a administré un produit pharmaceutique ou appliqué une intervention médicale, n'ayant pas nécessairement de lien de causalité avec le traitement, contrairement à l'effet indésirable. Un événement indésirable peut donc être un signe défavorable et imprévu (y compris un résultat de laboratoire anormal), un symptôme ou une maladie associé à l'utilisation du produit médical de recherche. **Ils sont classés selon leur aspect attendu et inattendu.**

Événement indésirable grave

Un événement indésirable grave (EIG) est défini comme un événement fâcheux d'ordre clinique et/ou biologique survenant à des doses thérapeutiques et qui est responsable des effets suivants :

- entraîne le décès du sujet ;
- met sa vie en danger ;
- nécessite son hospitalisation ou la prolongation de son hospitalisation ;
- entraîne une invalidité/une incapacité permanente ou importante ou se traduit par une anomalie/malformation congénitale.

Événement indésirable grave attendu

Un événement indésirable grave attendu (EIG-A) est défini comme un événement déjà mentionné dans la version la plus récente de la « brochure investigateur » ou dans le résumé des caractéristiques du produit (RCP) pour les médicaments ayant déjà une autorisation de mise sur le marché (AMM). Cette définition s'applique également au médicament de l'essai lorsqu'il est administré pour une même population hors indication de l'AMM.

Événement indésirable grave inattendu

Un événement indésirable grave inattendu (EIG-I ou SUSAR, *Suspected Unexpected Serious Adverse Reaction*) est défini comme un événement non mentionné ou différent par sa nature, son intensité, son évolution par rapport à la « brochure investigateur » ou au RCP pour les médicaments ayant une AMM.

Toxicité et effets indésirables

Dans les essais de phase I, l'évaluation de la toxicité et des effets secondaires est le critère principal pour la recherche de la dose recommandée en phase II. L'information recueillie dans ces essais est alors très détaillée. Dans les essais de phase II et III, la toxicité aiguë est évaluée pendant l'administration des traitements, durant l'intercure et souvent jusqu'à 90 jours après la dernière administration. Dans les essais de phase II, l'information recueillie sur les toxicités aiguës est souvent très détaillée. Dans les essais de phase III, l'information recueillie sur les toxicités aiguës est souvent limitée aux effets indésirables graves qu'ils soient attendus ou inattendus. Les toxicités chroniques ou complications tardives sont souvent évaluées quand les patients sont suivis plusieurs années après la fin du traitement, le plus souvent à partir du 3e ou 6e mois après la dernière administration.

À la conception de chaque étude, le recueil des toxicités est adapté à la pathologie et aux traitements administrés. Le rythme de suivi est également planifié. Les effets sont recueillis avant l'administration du traitement à l'étude *(baseline)* afin de connaître l'état initial du patient, puis au cours du traitement selon un rythme préétabli, identique dans les différents bras de traitement dans le cas d'un essai à plusieurs groupes.

Les toxicités, de manière générale, qu'elles soient liées à la chimiothérapie ou à la radiothérapie, provoquent des modifications cliniques et/ou biologiques. Dans le cas de la radiothérapie, elles sont provoquées par un surdosage sur des tissus sains. Ce sont les contraintes de la dosimétrie sur les organes à risque qui sont les éléments les plus critiques dans la conduite d'une irradiation : les altérations somatiques, fonctionnelles et structurales peuvent survenir durant le traitement lui-même (effets aigus), mais les effets tardifs peuvent se manifester des mois, voire des années, après la disparition des effets aigus et peuvent s'aggraver avec le temps. La meilleure stratégie thérapeutique est celle qui requiert non seulement une disparition complète de la tumeur, mais aussi l'obtention de dégâts aussi minimes que possible au niveau des tissus sains environnants.

Selon la période d'observation, la toxicité sera définie comme aiguë ou tardive.

Effets aigus

La toxicité est définie comme aiguë si les événements indésirables sont observés au cours du traitement puis dans les 3 mois après la fin du traitement. Les effets secondaires sont maintenant codés de manière homogène selon les échelles du NCI-CTC *(National Cancer Institute-Common Toxicity Criteria)* [1].

Effets tardifs

La toxicité sera définie comme tardive si les événements indésirables sont observés au-delà de 3 mois après la fin du traitement.

Les effets tardifs (ou complications) sont souvent utilisés comme un des critères principaux dans les études évaluant des traitements de radiothérapie, mais ils peuvent aussi bien concerner la chimiothérapie que la chirurgie. En radiothérapie, il est bien connu que la radiobiologie (doses totales, dose par fraction, débit de dose, etc.) influence de manière différente les effets tardifs et les effets aigus. La tolérance des tissus normaux à la radiothérapie reste le facteur limitant à la délivrance d'une dose tumoricide.

La préservation de la qualité des tissus sains est une préoccupation importante pour les cliniciens avec une prise en compte du risque de morbidité qui fait partie intégrante de la prise en charge quotidienne et du suivi des patients. Bien que la radiothérapie ait fait beaucoup de progrès au niveau des complications, il apparaît néanmoins nécessaire d'utiliser des échelles permettant d'évaluer les toxicités tardives et de suivre les patients régulièrement pour observer ces effets.

Ce fait permet de comprendre la nécessité ressentie de mettre en place un système d'évaluation précis et reproductible des effets tardifs. C'est particulièrement le cas de la radiothérapie, où plusieurs échelles, plus ou moins détaillées, ont vu le jour, chacune essayant de s'adapter à des localisations différentes.

Les échelles disponibles

Quelle que soit l'étape de développement d'un médicament, il est important de choisir l'échelle la plus appropriée pour en évaluer la toxicité afin d'en déterminer sa tolérance, car il en existe plusieurs, validées ou non, plus ou moins utilisées.

Le *Radiation Therapy Oncology Group* (RTOG) en Amérique du Nord a été un des premiers à développer une échelle pour l'évaluation des effets tardifs dus à la radiothérapie, qui a donné lieu à un consensus en commun avec l'*European Organization for Research and Treatment of Cancer* (EORTC) (1995). L'échelle LENT-SOMA (*Late Effects on Normal Tissue* – subjectivité, objectivité, management et analyse) a été développée pour l'évaluation des effets tardifs des tissus normaux par l'EORTC en Europe, mais ces échelles sont difficiles à appliquer et peu utilisés dans leur ensemble [2, 3].

Le choix de l'échelle va dépendre des effets connus du traitement à l'étude ainsi que de la localisation. Par exemple, l'échelle habituelle de toxicité dans une étude sur un nouveau traitement qui présente des effets gastro-intestinaux attendus peut être supplémentée avec une comptabilité journalière des événements indésirables.

L'échelle de DAHANCA, adaptée pour détecter de manière objective des toxicités des traitements du cancer de la tête et du cou, est utilisée en Danemark mais, comme beaucoup de questionnaires, cette échelle n'a pas été validée. Elle a montré une faible sensibilité pour détecter des plaintes subjectives [4]. D'autres échelles, comme le glossaire franco-italien développé spécifiquement pour le cancer gynécologique afin de détecter des effets tardifs [5], ont été peu utilisées.

Une comparaison entre échelles permet de les évaluer avant de choisir celle qui semble la plus adaptée. Par exemple, les résultats des toxicités de l'essai GORTEC 94-01 ont comparé trois échelles et ont conclu que l'échelle LENT/SOMA semblait la plus précise mais que la plupart des scores n'étaient pas concordants avec ceux des autres échelles. Ce constat confirme la nécessité d'utiliser des échelles communes pour évaluer les effets tardifs dans les essais cliniques [6].

Un autre exemple dans le cancer du poumon chez des patients traités par un traitement d'irradiation à visée curative a permis de conclure que l'évaluation de ces toxicités dépend de l'échelle utilisée et qu'il faut être prudent dans les interprétations. À titre d'illustration, l'échelle RTOG/EORTC a détecté 23 % de toxicités de grade 3 alors que l'échelle NCI-CTC, qui ne prend pas en compte les anomalies radiologiques, n'en a détecté aucune [7].

Depuis la version 3 de l'échelle NCI-CTC (CTCAE v3.0, 2002), la cotation permet d'évaluer sur les mêmes référentiels les toxicités aiguës et tardives en concernant aussi bien les toxicités radio-induites que chimio-induites.

La version 4 de l'échelle NCI-CTC (CTCAE v4.0, 2009) est très différente de la version 3 et est basée sur les catégories par classe d'organe (SOC pour *System Organ Class*) du MedDRA. La façon la plus simple et la plus rapide de trouver un terme est d'utiliser la version électronique.

Le MedDRA correspond à un code de la pathologie ou de la condition médicale étudiée dans l'essai clinique, établi selon le dictionnaire MedDRA® (*Medical Dictionary for Regulatory Activities*), créé à l'initiative de l'ICH *(International Conference on Harmonization of Technical Requirements of Registration of Pharmaceuticals for Human use)*.

Après avoir défini l'outil d'évaluation de la toxicité (choix de l'échelle de cotation), il est nécessaire d'en définir le rythme d'évaluation.

La toxicité hématologique, par exemple, si elle est évaluée uniquement au début de chaque cycle de chimiothérapie, ne serait pas suffisante pour estimer le nadir (la valeur la plus basse d'une mesure au cours du temps), ni la durée de l'aplasie. Les études planifiées spécifiquement pour évaluer les effets des facteurs de croissance sur la toxicité hématologique devraient être plus contraignantes dans la définition du rythme d'évaluation. Les études où la fréquence de l'évaluation diffère entre les groupes de traitements peuvent produire des comparaisons biaisées.

Cette même difficulté est présente pour l'évaluation des toxicités tardives, car le rythme de surveillance est de plus en plus espacé dans le temps, et il sera plus difficile de repérer des complications de faible grade que celles de grade 3 ou 4 ou de renseigner leur durée.

Utilisation des données de toxicité et exemples

Critère de jugement d'un essai

Les données de toxicité peuvent être utilisées comme critère de jugement dans les essais cliniques notamment pour les phases II et III.

Essais de phase II

Le plan d'expérience développé par Bryant et Day permet de planifier un essai de phase II avec un critère combiné d'efficacité et de toxicité [8]. Pour cela, il est nécessaire de fixer *a priori* le seuil d'efficacité que l'on souhaite atteindre et le seuil de toxicité à ne pas dépasser. Ce plan permet de statuer en deux étapes sur une règle d'arrêt reposant à la fois sur le critère d'efficacité et sur le critère de toxicité. Si au moins un des deux critères n'est pas respecté, l'essai est arrêté (*cf.* chapitre V.2, page 281).

D'autres plans inspirés du monde bayésien ont été développés et sont présentés dans d'autres chapitres (*cf.* chapitres I.6 « Statistiques bayésiennes », page 51 et V.2, page 281). Ces méthodes sont très intéressantes mais complexes car elles nécessitent des calculs intensifs avant leur mise en œuvre.

Arrêt précoce

Dans le cadre des essais de phase III, des méthodes statistiques basées sur un monitorage séquentiel d'EIG pouvant arriver relativement tôt permettent un arrêt des inclusions tout en préservant le risque de première espèce [9]. La méthode développée définit des règles d'arrêt après la survenue de chaque EIG en comparant le nombre de patients inclus au nombre de patients satisfaisant le maximum d'EIG. Elle est surtout intéressante à appliquer par le promoteur d'un essai multicentrique où chaque centre n'a pas forcément une vision globale de la survenue de l'ensemble des EIG.

À titre d'exemple, on fixe *a priori* le nombre total de sujets nécessaires planifié dans l'essai sur le critère principal, le taux d'EIG à ne pas dépasser, et si les x premiers patients présentent au moins un EIG parmi les y premiers patients inclus dans l'essai, les inclusions sont suspendues en attendant une décision du comité indépendant de surveillance (*cf.* chapitre VI.3 « Les comités indépendants de surveillance des essais thérapeutiques : rôle et modalités de fonctionnement », page 354), sinon les inclusions peuvent continuer.

Rapports

En parallèle des planifications et du suivi de l'essai clinique, il existe des exigences réglementaires permettant d'assurer la sécurité des personnes par le suivi de la totalité des EIG observés, qui doivent être déclarés régulièrement.

Rapport annuel de sécurité

Un rapport annuel de sécurité est à envoyer par le promoteur d'essais cliniques une fois par an ou sur demande de l'Agence française de sécurité sanitaire des produits de santé (Afssaps) [10].

Son objectif est d'évaluer de façon annuelle la sécurité des personnes participant à un essai clinique.

Ce rapport est une analyse globale concise, pertinente de toute information de sécurité disponible concernant les essais et les traitements expérimentaux pendant la période considérée. Ce n'est pas une demande d'amendement ou un résumé du rapport de fin d'essai.

Il est constitué des trois parties suivantes :

- l'analyse de la sécurité des personnes dans l'essai : description et analyse des nouvelles données pertinentes de sécurité relatives à la sécurité du traitement expérimental, analyse critique de leur impact sur la sécurité des participants, le rapport bénéfice/risque, les mesures prises pour minimiser les risques ;
- la liste de tous les EIG dans l'essai concerné (EIG attendus, inattendus depuis le début de l'essai) selon différentes sections (traitement expérimental, traitement comparateur) et des rubriques principales pour chaque cas (ref. essai, ref. cas, âge, sexe, pays, effet, traitement étudié, caractère attendu, causalité) ;
- et des tableaux de synthèse récapitulant les termes relatifs aux EIG, depuis le début de l'essai, permettant une vision globale, spécifiant pour chaque système organe et pour chaque terme d'effet indésirable le nombre d'EIG regroupé selon le bras de traitement en précisant leur caractère attendu ou non.

En cas d'essai multiple, un seul rapport annuel de sécurité est nécessaire avec une analyse globale du profil de sécurité du médicament expérimental testé jusqu'à la date de clôture du dernier essai dans l'Union européenne (60 jours après la clôture, 90 jours après la fin de l'essai en cas de 1re administration à l'homme).

Un rapport annuel de sécurité est également à transmettre auprès du Comité de protection des personnes (CPP) identique à celui de l'Afssaps.

Rapport d'essai clinique

Dans le rapport d'essai clinique, les données de l'évaluation de la toxicité seront décrites pour l'ensemble des patients traités selon les statistiques usuelles.

Pour chaque type de toxicité, le nombre de chacun des grades observés sera décrit par cycle (% grade sur le nombre total de cycles administrés) et par patient (% de patients présentant une toxicité spécifique selon le grade maximal observé au cours du traitement).

De plus, seront décrits les grades 3 ou plus de la toxicité observée par cycle (sur le nombre total de cycles administrés) et par patient (grade maximal observé au cours du traitement).

Cette mesure sur l'ensemble des cycles ne tient pas compte du fait que le même patient peut présenter la même toxicité à chaque cycle. Elle considérera un patient qui présentera 4 toxicités différentes de grade 3 de la même façon que 4 patients qui présenteront une seule toxicité de grade 3.

Les toxicités seront décrites selon les catégories aiguës et tardives s'il y a lieu.

Les effets toxiques sont généralement regroupés par type de toxicité et présentés selon les toxicités hématologique et non hématologique.

Remarque

Le test statistique du chi-2 comparant les grades de toxicité entre deux groupes n'est pas toujours bien adapté car ce test ignore la nature ordonnée des grades de toxicité. Il est préférable d'utiliser le test de Kruskal-Wallis (*cf.* chapitre I.4 « Choix du bon test statistique », page 28).

Des analyses plus complexes peuvent être produites comme, par exemple, des corrélations entre les paramètres initiaux et la survenue des événements toxiques, des évaluations longitudinales de la survenue des événements toxiques par l'utilisation de modèles mixtes qui permettent de mesurer l'effet temps dans l'évolution de la toxicité, des méthodes d'analyse de survie permettant d'évaluer le délai d'apparition ou de disparition de certains types de complications.

Brochure investigateur et profil de tolérance pour autorisation de mise sur le marché

À la mise en place d'un essai clinique, les RCP sont clairement identifiés pour chacun des produits testés en annexe du protocole et dans la « brochure investigateur » mise à disposition lors de la mise en place de l'essai clinique.

Les adaptations de doses à réaliser en cas de toxicités observées au cours de l'administration du traitement étudié sont alors décrites précisément dans le protocole afin de connaître la conduite à tenir en cas d'une manifestation de toxicité.

Le profil de tolérance ainsi observé au cours d'un essai peut être comparé aux toxicités décrites dans les RCP d'un médicament, et des études de suivi post-AMM permettent de valider ce profil de tolérance observé au cours du développement thérapeutique de phase I, II et III.

À retenir
- La notion de toxicité aiguë et tardive
- L'échelle de cotation unique et les échelles adaptées au type de toxicité spécifiques
- Le type de rapport de données de toxicité
- Les méthodes de prise en compte de la toxicité dans la planification des essais de phase II et III

Références

1. Common Terminology Criteria for Adverse Events (CTCAE) and Common Toxicity Criteria (CTC) CTC-AE v4.0: http://ctep.cancer.gov/protocolDevelopment/electronic_applications/docs/ctcaev4.pdf

2. Pavy JJ, Denekamp J, Letschert J, *et al*. EORTC Late Effects Working Group. Late effects toxicity scoring: The SOMA scale. *Int J Radiat Oncol Biol Phys* 1995 ; 31 (5) : 1043-7.

3. Cox JD, Stetz J, Pajak TF. Toxicity criteria of the Radiation Therapy Oncology Group (RTOG) and the European Organization for Research and Treatment of Cancer (EORTC). *Int J Radiat Oncol Biol Phys* 1995 ; 31 (5) : 1341-6.

4. Jensen K, Bonde Jensen A, Grau C. The relationship between observer-based toxicity scoring and patient assessed symptom severity after treatment for head and neck cancer. A correlative cross sectional study of the DAHANCA toxicity scoring system and the EORTC quality of life questionnaires. *Radiother Oncol* 2006 ; 78 (3) : 298-305.

5. Chassagne D, Sismondi P, Horiot JC, *et al.* A glossary for reporting complications of treatment in gynecological cancers. *Radiother Oncol* 1993 ; 26 (3) : 195-202.

6. Denis F, Garaud P, Bardet E, *et al*. Late toxicity results of the GORTEC 94-01 randomized trial comparing radiotherapy with concomitant radiochemotherapy for advanced-stage oropharynx carcinoma: Comparison of LENT/SOMA, RTOG/EORTC, and NCI-CTC scoring systems. *Int J Radiat Oncol Biol Phys* 2003 ; 55 (1) : 93-8.

7. Faria SL, Aslani M, Tafazoli FS, Souhami L, Freeman CR. The challenge of scoring radiation-induced lung toxicity. *Clinical Oncology* 2009 ; 21 (5) : 371-5.

8. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995 ; 51 (4) : 1372-83.

9. Kramar A, Bascoul-Mollevi C. Early stopping rules in clinical trials based on sequential monitoring of serious adverse events. *Med Decis Making* 2009 ; 29 : 343-50.

10. Afssaps. *Rapport annuel de sécurité :* http://www.afssaps.fr/var/afssaps_site/storage/original/application/efb035c841985d31ff8e104ca88b8a95.pdf

La survie comme critère de jugement

S. Mathoulin-Pélissier, S. Gourgou-Bourgade, F. Bonnetain

Les critères cliniques sont les critères de jugement qui correspondent aux objectifs thérapeutiques d'un traitement. Leur nature varie en fonction de l'objectif thérapeutique. Le critère principal de jugement *(primary endpoint)* doit être la variable la plus pertinente sur le plan clinique pour évaluer l'efficacité d'un traitement. Un point crucial concerne la définition la plus précise et la plus objective possible de ce critère. La justification du choix du critère et les méthodes retenues pour sa mesure (test biologique, méthodes cliniques, questionnaires, etc.) doivent être parfaitement décrites dans le protocole. Les critères secondaires concernent les autres mesures d'efficacité ou de tolérance du traitement testé.

> Le critère de jugement est la traduction d'un événement d'intérêt en une variable permettant de répondre à la question d'un projet de recherche. Ainsi, objectif principal de l'étude, événement d'intérêt et critère de jugement sont obligatoirement liés dans la construction de la recherche. Par exemple, si on s'intéresse à l'efficacité d'un traitement cytotoxique pour un cancer, le critère de jugement sera la survie globale et l'événement d'intérêt peut être le décès quelle qu'en soit la cause.

L'objectif de toute nouvelle intervention thérapeutique en cancérologie est d'améliorer la survie et/ou la qualité de vie des patients atteints d'un cancer. Ainsi, la survie globale demeure un critère de jugement de choix dans les essais de phase III évaluant un nouveau traitement ou une nouvelle stratégie. Seulement, l'analyse de ce critère de jugement ne peut se faire que lorsqu'un nombre important de décès a eu lieu. Cela est généralement assez rapide dans certains types tumoraux à mauvais pronostic comme les cancers du pancréas ou les hépatocarcinomes, mais beaucoup moins lorsqu'il s'agit d'évaluer un nouveau traitement en situation adjuvante dans le cancer du sein par exemple. C'est l'une des raisons pour lesquelles l'utilisation d'autres critères de jugement de survie comme la survie sans récidive, la survie sans progression est devenue de plus en plus fréquente dans les essais cliniques. Ces derniers critères sont des critères composés de plusieurs événements *(composite endpoints)*, et ils sont souvent utilisés comme des critères de substitution *(surrogate endpoints)*. Nous allons expliquer ces notions dans ce chapitre qui n'abordera pas les notions de survie relative (pour plus de détails sur cette technique d'analyse : [1]).

Il existe en fait des conditions assez intuitives pour qu'un critère de jugement soit utile :

- la pertinence clinique : ce qu'il mesure fait une réelle différence dans la vie du patient ou sa prise en charge (par exemple, l'efficacité d'une nouvelle molécule cytotoxique et la survie augmentée ou la diminution de récidives). Ce point soulève des questions quant à la relation démontrée entre critères intermédiaires (progression, récidives) et critère terminal (décès) et quant à l'acceptabilité du résultat pour la pratique clinique ;
- la validité : l'événement d'intérêt doit être mesuré sans biais ;
- la fiabilité : la mesure de l'événement d'intérêt doit être répétable et reproductible (par exemple la mesure d'un marqueur biologique comme les PSA dans l'évaluation de l'efficacité d'un traitement pour les cancers de la prostate ou la progression clinique et radiologique pour de nombreux cancers).

Les délais de survie comme critères de jugement

La durée de survie constitue le critère majeur d'évaluation du traitement de nombreux cancers. Elle est étudiée grâce à la constitution de courbes de survie estimant, à partir de la découverte de la maladie (ou tout autre événement choisi comme origine), la probabilité pour les sujets d'être vivants lors de délais successifs.

Un point crucial concerne la définition la plus précise et la plus objective possible de ce critère. Par exemple, dans le cas où l'on étudie l'efficacité d'un traitement sur la mortalité, il faudra préciser si le critère principal consiste en la proportion de patients vivants à une date fixée après la mise sous traitement ou en la durée de survie après la mise sous traitement. Cela est d'autant plus important que le choix de ce critère conditionne le nombre de sujets à inclure dans l'essai.

La justification du choix du critère et les méthodes retenues pour sa mesure (test biologique, méthodes cliniques, radiologiques, suivi et dates de suivi) devront être parfaitement décrites dans le protocole. Les critères secondaires concernent les autres mesures d'efficacité ou de tolérance du traitement testé. Ceux-ci doivent également être définis (ainsi que leur mode de mesure) dans le protocole.

Dans ce chapitre, nous expliquerons les vocabulaires de survie (survie globale, survie sans maladie, survie sans récidive, etc.) les plus employés en explicitant les pièges de leur utilisation et en nous appuyant sur des exemples tirés d'essais cliniques randomisés, essais cliniques qui sont le *gold standard* (étalon) pour évaluer l'effet d'un traitement.

Des éléments minimaux de définition par patient sont indispensables à connaître pour ce chapitre ; ils seront repris et détaillés dans le chapitre III.1 « Données de survie » (*cf.* page 129) : un événement ou plusieurs (par exemple, décès toute cause, la récidive, la progression), un patient perdu de vue, un patient censuré, des dates… Ce sont là des bases indispensables.

La survie globale

Définition

Tout critère de survie sera défini par trois paramètres : la date d'origine, le délai entre l'origine et l'événement à observer et si l'événement n'est pas observé à la date de dernière nouvelle ou à la date point de l'analyse : la censure.

La survie globale *(overall survival)* est le critère de survie de référence pour l'évaluation des essais de phase III où l'on veut démontrer les effets à long terme d'un traitement ayant démontré une activité antitumorale en phase II. Il s'agit du paramètre ultime que l'on cherche à améliorer dans tous les essais cliniques en cancérologie, particulièrement en situation de traitement principal ou adjuvant.

La survie globale est définie plus précisément comme le temps entre l'entrée dans l'essai clinique (randomisation) ou le début du traitement (date d'origine) jusqu'au décès quelle qu'en soit la cause (événement). Les patients vivants seront censurés aux dernières nouvelles ou à la date de point (censure). La censure indique que le patient n'a pas eu l'événement d'intérêt jusqu'à la date de dernière nouvelle ou la date de point.

La survie globale est un critère objectif facile à recueillir mais qui nécessite un suivi particulier afin de ne pas avoir trop de perdus de vue (patients dont on n'a plus de nouvelles et pour lesquels cette information n'est peut-être pas aléatoire mais le reflet d'un effet du traitement). Contrairement aux autres critères de survie qui peuvent être utilisés, la survie globale ne pose pas de problème de définition ni de problème d'interprétation clinique. La survie globale est généralement utilisée comme critère de jugement principal pour les cancers à mauvais pronostic en situation adjuvante car c'est avant tout ce bénéfice clinique que l'on va chercher à améliorer.

Un critère secondaire parfois utilisé, surtout chez les populations les plus âgées, sera la survie spécifique. Elle se différencie de la survie globale car elle prend en compte comme événement uniquement les décès d'une cause spécifique (lié directement au cancer). Cependant, il sera parfois difficile de ne pas imputer le décès au cancer si le décès est survenu suite à une toxicité induite par le traitement du cancer. Toute la difficulté réside dans le choix de l'attribution de la cause spécifique à un certain moment dans l'histoire de la maladie. L'idéal dans un essai clinique de phase III est que la cause de décès soit évaluée à l'aveugle par un comité indépendant de décision de l'attribution de la cause de décès au vu de toutes des données du patient.

Précautions

En évaluant la survie globale, il existe deux limites potentielles en cas de longue survie – un nombre important de malades peuvent être perdus de vue – et lors de traitements de rattrapage qui peuvent être donnés en cas d'échec du traitement évalué.

Un suivi trop long entraîne un risque de perdus de vue et donc un risque de perte d'information non lié au hasard (biais d'information).

Dans l'autre cas, le traitement de rattrapage peut compenser l'échec initial (survenue d'un événement plus précoce : rechute ou progression de la maladie), et l'interprétation de l'effet du traitement initial sur la survie globale sera biaisée. En effet, la mesure de la survie globale sera le reflet de l'effet de la stratégie thérapeutique globale (les deux traitements) depuis la date d'origine jusqu'à l'obtention de l'événement décès (ou sa non-obtention) prenant en compte le traitement initialement testé et celui de rattrapage.

Dans les situations où l'observation de l'événement « décès » serait trop lointaine, des critères observables plus précocement semblent intéressants, afin d'obtenir une réponse plus rapide sur l'efficacité. Ces critères observés plus précocement peuvent être la progression de la maladie, la rechute locale et/ou métastatique, un second cancer, des complications liées au traitement ou la qualité de vie relative à la santé. Afin d'être utilisés comme critère principal, ces critères plus précoces de l'efficacité doivent avoir été validés comme étant prédictifs de l'effet sur la survie globale ou être en eux-mêmes des critères cliniques. Il s'agit ici des notions relatives aux critères intermédiaires potentiellement substitutifs de la survie globale qui sont expliqués dans le chapitre II.5 (« Critères de substitution », page 113).

Exemples dans les essais cliniques

La survie globale comme critère principal est généralement utilisée dans les essais cliniques pour les cancers de mauvais pronostic car l'histoire de la maladie fait qu'il y aura un intervalle très court entre le diagnostic du cancer, la rechute et la survenue du décès. Un autre élément explicatif de cette utilisation est souvent l'absence d'alternative dans le traitement de la rechute – donc la survie sans rechute n'est pas non plus pertinente –, par exemple, pour le cancer du pancréas dont la médiane de survie globale en 1re ligne métastatique est de l'ordre de 11 mois [2]. L'objectif d'un nouveau traitement sera d'améliorer ce temps de survie. Pour les cancers de meilleur pronostic comme pour le cancer du sein, des critères d'évaluation plus précoces sont envisagés afin de répondre à la question dans un délai raisonnable et avec un nombre de patients qui pourra être moins important. En effet, le cancer du sein non métastatique est devenu une maladie qui, aujourd'hui, peut être traitée par de nombreuses lignes de mono- ou polychimiothérapies associées ou non à des thérapies ciblées. Ces associations font l'objet de nombreux essais thérapeutiques dont le critère principal utilisé est généralement la survie sans progression (PFS pour *Progression Free Survival*) plutôt que la survie globale afin d'évaluer l'efficacité de la séquence thérapeutique expérimentale et non la prise en charge globale. Souvent, les essais montrent une amélioration de la PFS à chacune des « lignes » thérapeutiques (séquence de traitements différents) sans en améliorer systématiquement la survie globale. Ainsi, l'adjonction de trastuzumab a nettement permis d'améliorer la médiane de PFS depuis 10 ans mais le recul pour estimer la survie globale est encore insuffisant [3].

Il en est de même pour le cancer du sein non métastatique, en situation adjuvante, où les survies ont été nettement améliorées. Il faudrait des échantillons de population très importants avec un suivi très long pour démontrer un gain de survie globale cliniquement intéressant. De plus, la prise en charge des patientes est multiple et plusieurs lignes de traitement sont observées avant le décès. Après chirurgie, chez les patientes avec des récepteurs hormonaux positifs (RH+), un

traitement hormonal au long cours sera proposé. Afin d'en évaluer l'effet sur la survie globale qui est voisine de 85 % à 5 ans, il faudrait un gain conséquent pour une taille d'échantillon raisonnable. Sinon, avec un gain de 2-3 %, le nombre de sujets à inclure est considérable, justifiant de grands essais obligatoirement internationaux. De plus, la prise en charge de la maladie jusqu'au décès fait intervenir de nombreuses « lignes » de traitement – différentes chimiothérapies se succèdent et les pratiques peuvent varier après l'administration de la chimiothérapie que l'on veut en fait évaluer –, et il n'est donc pas possible de conclure à l'effet propre du traitement adjuvant mis en œuvre. C'est pour ces raisons qu'il est recommandé au niveau international d'évaluer ce type de traitement par un critère observé plus précocement et de préférence dans la fenêtre temporelle de ligne de traitement évaluée comme la survie sans maladie (DFS pour *Disease Free Survival*) à 5 ans. En ayant un effet sur cette DFS, on fait l'hypothèse qu'en retardant la rechute, on aurait un effet sur la survie globale. Des échantillons de population importants sont cependant nécessaires pour observer un gain clinique, et ces essais sont le plus souvent internationaux [4]. En fait, dans ces exemples, nous avons abordé d'autres termes de survie (PFS, DFS) qui sont des critères de survie composés de plusieurs événements (critères « composites »). Ce caractère composite renvoie à l'importance de la définition des événements.

Critères composites[1] et survie

L'histoire de la maladie fait qu'il existe de nombreux autres événements carcinologiques avant le décès. Dans certaines situations thérapeutiques, la survenue du décès se fera après un long délai et il sera nécessaire d'améliorer chaque intervalle de temps entre les différents événements petit à petit pour au final espérer améliorer la survie globale. C'est pourquoi il existe de nombreux autres critères évalués plus précocement qui sont le plus souvent des critères composites.

Un critère composé ou « composite » *(composite endpoint)* est un critère qui prend en considération simultanément plusieurs événements d'intérêt [5]. Le critère « événements cancer mortels ou non mortels » est un critère composite formé du regroupement des événements cancer non mortels (rechute, métastase) et des décès liés au cancer. Un patient présente le critère composite à partir du moment où il a développé l'un de ces événements. En cancérologie, la majorité des critères de survie autres que la survie globale sont actuellement des critères composites *(tableau I)*.

> Un critère de jugement composite est la prise en compte, dans une même variable, de plusieurs événements d'intérêt mesurés chacun au niveau d'un individu (une combinaison de plusieurs événements).
> Pour le cancer du sein, l'efficacité d'un nouveau traitement cytotoxique peut se mesurer par plusieurs critères : le nombre de décès, de métastases et de rechutes locorégionales évités ; ce critère de jugement, appelé aussi survie sans maladie, est un critère de survie composite.

1. Différent d'un score composite, non abordé ici.

La survenue successive de plusieurs des composantes du critère « composite » ne donne lieu qu'à une seule occurrence du critère composite. En général, on sélectionne celui qui survient en premier. Par exemple, un patient présentant successivement une récidive, puis décédant du cancer ne sera compté qu'une seule fois pour le critère composite. De ce fait, le nombre d'occurrences du critère composite n'est en général pas la somme des occurrences de chacune de ses composantes. Par exemple, la survie sans progression mesure le délai jusqu'au critère composite défini comme le 1[er] événement observé entre la progression de la maladie ou le décès ; ce critère composite est alors appliqué et utilisé dans les essais cliniques en tant que critère principal ou secondaire, pour différentes localisations, par exemple dans le cancer colorectal en situation avancée [6].

Tableau I. Les survies : définitions de critères composites tels que survie sans progression et survie globale.

Événement		**Critères**		**Information sur l'évolution de la maladie ou l'effet du traitement**
		Survie sans progression	**Survie globale**	
Progression : augmentation tumorale		E	I	Oui
Progression : nouvelle localisation		E	I	Oui
Décès	Lié au même cancer	E	E	Oui
	Lié à un autre cancer	E	E	Possible
	Lié à la toxicité du traitement	E	E	Oui
	Lié à d'autres causes	E	E	Possible
Vivants et sans progression au dernier suivi		C	C	Non applicable

*E : événement, C : censure, I : ignoré.

L'utilisation des critères composites présente plusieurs intérêts :
- augmenter la puissance statistique ;
- mesurer directement le rapport bénéfice/risque (regroupement de plusieurs événements délétères mesurant un seul effet négatif) ;
- ou regrouper des équivalents du même phénomène clinique.

Cette dernière situation est sûrement la justification principale de leur utilisation en cancérologie pour la plupart des cancers ; en effet, récidive, progression et décès sont les événements clés de la prise en charge thérapeutique des cancers. La survenue d'un de ces événements peut ainsi être assimilée à un échec thérapeutique et ces éléments être jugés « équivalents » dans la finalité de l'évaluation : optimiser la prise en charge thérapeutique afin d'assurer un bénéfice clinique au patient.

Le deuxième intérêt recherché est d'augmenter la probabilité de démontrer un effet. Étant donné que le nombre d'événements est plus grand, les critères de jugement « composites » augmentent la puissance statistique pour démontrer l'éventuel effet du traitement expérimental. De fait, le nombre de sujets nécessaires est moindre et/ou la durée d'étude est plus courte avec un critère composite qu'avec une seule de ses composantes. Cette justification est aujourd'hui particulièrement pertinente pour des cancers de bon pronostic afin de disposer plus rapidement des résultats. Ainsi, par exemple, un ensemble d'événements peut être comptabilisé dans la mesure d'un délai de « survie sans événements », chacun des événements devant être bien sûr défini et pouvant varier en fonction du type de cancer et de son stade.

Un autre intérêt serait de mesurer plus fidèlement le rapport bénéfice/risque du traitement en composant les événements des critères composites par des événements négatifs comme le décès du cancer mais aussi les complications graves liées au traitement reçu comme les 2es cancers ou les décès liés à la toxicité du traitement, par exemple. Cette approche donne cependant le même poids aux événements prévenus et aux événements induits ; cela est une hypothèse forte nécessitant des précautions. En effet, de tels critères composites posent fréquemment des problèmes d'interprétation, particulièrement quand les événements pris en compte ont des gravités cliniques différentes et que l'effet du traitement est différent sur chacun des événements. Enfin, il a été récemment mis en avant une hétérogénéité des définitions d'un même critère de survie à partir d'une analyse de la littérature [7].

Critères intermédiaires et de substitution

> Les critères de jugement intermédiaires *(intermediate endpoint)* ne correspondent pas directement à un objectif thérapeutique et documentent plutôt les mécanismes d'actions du traitement. Il s'agit le plus souvent de paramètres biologiques ou physiologiques. Ils sont communément dénommés biomarqueurs [8].

En cancérologie, les critères intermédiaires les plus utilisés sont la réponse tumorale, la survie sans progression ou la survie sans maladie comme des marqueurs d'une activité du traitement sur la tumeur. Ils sont et/ou devraient être le plus souvent utilisés dans les essais de phase II. Leur utilisation dans des essais de phase III peut se faire mais reste sujette à des extrapolations sur le bénéfice clinique puisque ces critères intermédiaires ne sont pas toujours des critères de substitution validés d'un critère clinique tel que la survie globale.

En effet, quels seraient les bénéfices pour le patient d'un traitement améliorant la survie sans progression mais sans amélioration de la survie globale et/ou de la qualité de vie relative à la santé ?

> Un critère de substitution est un critère intermédiaire (biomarqueur) utilisé à la place d'un critère clinique en tant que critère de jugement principal d'un essai. Un changement induit par le traitement sur le critère de substitution est supposé prédire un changement pour le critère clinique final.

Les critères de survie composites sont ainsi souvent utilisés, à tort ou à raison pour certains, comme des critères de substitution afin de remplacer le critère final clinique de survie globale ou de qualité de vie. Pour utiliser ces critères en tant que critère de jugement principal, il existe aujourd'hui des méthodes statistiques qui permettent de le valider en tant que critère de substitution préalablement. Un chapitre dans ce livre (chapitre II.5, page 113) est consacré à ces méthodes. Par ailleurs, les agences européennes et américaines donnent des recommandations pour l'utilisation ou la non-utilisation de ces critères dans les essais de phase III en vue de l'obtention d'une mise sur le marché de produits de santé [9].

Les limites des critères de survie pour l'interprétation des résultats

Concernant le critère de survie globale, nous avons déjà abordé les limites à son utilisation avec la difficulté liée à l'évaluation non pas d'un traitement mais de plusieurs si l'ensemble de la stratégie thérapeutique n'est pas randomisé. Pour les critères composites, il peut exister plusieurs limites qui se comprennent facilement : la dilution de l'effet (puisqu'il y a plusieurs événements) si les traitements n'ont pas des effets convergents pour les différents événements considérés ; le problème de la multiplicité des tests statistiques si la survie était testée par type d'événements (pour faire cela, il faudra au moins ajuster le risque alpha ou présenter les résultats comme secondaires et exploratoires). Enfin, deux autres dangers sont présents avec ces critères :

- la non-prise en compte de risques compétitifs : il s'agit de la combinaison d'événements fatals et non fatals ; un traitement responsable de plus de décès peut apparaître comme protecteur vis-à-vis des événements non fatals (récidives/progression) ;
- et la présence de données incomplètes : par exemple, l'arrêt du suivi après survenue d'un événement non fatal. Un chapitre complet lui est dédié (chapitre III.4 « Suivi et surveillance », page 181).

***Illustration :* la survie sans progression comme critère de jugement principal et les principaux pièges à éviter**

La progression, sa définition et sa mesure posent encore de nombreuses interrogations [10].

- Survie sans progression (SSP) et temps jusqu'à progression (TTP) sont deux critères de jugement qui diffèrent dans la mesure où les décès sans évidence de progression sont comptés comme des événements pour la SSP tandis qu'ils sont censurés pour le TTP. Dans le cas extrême où il y aurait beaucoup de décès toxiques sans mise en évidence de progression dans le bras expérimental d'un essai, le TTP pourrait paradoxalement être plus long dans ce dernier bras car ces derniers événements seraient censurés, alors que la survie globale serait moins bonne. Il est donc généralement plus approprié d'utiliser la SSP que le TTP.
- La définition de la progression : laquelle est le reflet de l'effet thérapeutique ? La progression clinique et/ou radiologique sur imagerie et/ou la progression des marqueurs tumoraux [11] ?
- En cas de progression radiologique, il est indispensable que l'imagerie soit revue par un comité de relecture indépendant afin de diminuer le risque de résultats biaisés. Même dans les essais en double aveugle, il est parfois facile de deviner dans quel bras de traitement est le patient en observant les toxicités. Lorsqu'il s'agit de progression clinique, le jugement est purement subjectif et peut donc aussi conduire à des résultats biaisés.
- Les évaluations tumorales clinique et radiologique doivent se faire à la même fréquence dans les deux bras de traitements. Une évaluation plus tardive d'une progression tumorale allongera de façon artificielle la SSP pour un patient donné. Mais cela pose aussi la question des intervalles de temps requis pour évaluer la progression : toutes les 4 ou 8 semaines ? Cet intervalle de temps est primordial car il peut engendrer un biais différentiel dans l'estimation de la SSP [12]. La SSP est, du fait du design des essais, systématiquement surestimée ; puisque le vrai temps de progression se situe dans l'intervalle de deux évaluations à temps fixe. En sus, du fait de signes cliniques évocateurs (douleur, fatigue), la tentation légitime d'avancer l'évaluation pour diagnostiquer la progression peut engendrer un biais différentiel.
- Il est fondamental que les patients continuent d'être suivis jusqu'à leur décès et qu'ils ne soient pas censurés et ce, même après mise en évidence d'une progression. La 1re raison pour cela est qu'une différence significative en SSP ne se traduit pas nécessairement en une différence en survie globale. La 2nde raison est que ces données de survie seront le seul moyen d'analyser ultérieurement la survie globale de façon adéquate afin de valider définitivement les résultats de l'essai.
- Les méthodes statistiques usuellement utilisées (courbe de survie de Kaplan-Meier) pour évaluer la SSP ne sont pas optimales puisqu'elles ne prennent pas en compte les risques compétitifs et les censures par intervalle.
- Quand la SSP est substitutive, la plus petite différence de SSP à observer pour espérer observer un gain cliniquement pertinent n'est pas établie de façon consensuelle.

Au total, ce sont donc souvent les notions de fiabilité et validité de ce critère de survie qui demandent à être affirmées de façon rigoureuse.

Les recommandations internationales

Consensus international : *Consort Statement*

L'énoncé CONSORT 2001 pour les essais cliniques randomisés est composé de 22 items et recommande dans l'item 6 dédié aux critères de jugement de définir clairement les critères objectifs primaires et secondaires et, lorsque c'est applicable, la méthode pour améliorer la qualité de la mesure du critère (observation multiple, formation des évaluateurs). La publication principale et sa récente mise à jour [13, 14] donnent pour chaque item un exemple et une explication du contenu recommandé par l'énoncé. Cependant, en dehors de spécifier qu'il faut définir complètement le critère, il n'y a pas de recommandation précise pour l'utilisation des critères de survie et surtout pour leur définition.

Les recommandations internationales : *International Conference on Harmonisation ± Efficacy* (ICH-E) [15]

Ces recommandations décrivent la façon de conduire et de rapporter les résultats des essais cliniques. Ces recommandations font l'objet des 20 documents suivants pour relater l'efficacité : *Clinical safety* (E1, E2A-E2F) ; *Clinical Study Reports* (E3) ; *Dose-Response Studies* (E4) ; *Ethnic Factors* (E5) ; *Good Clinical Practice* (E6) ; *Clinical Trials* (E7, E8, E9, E10, E11) ; ICH E7 *Studies in Support of Special Populations : Geriatrics* ; ICH E8 *General Considerations for Clinical Trials* ; ICH E9 *Statistical Principles for Clinical Trials* ; ICH E10 *Choice of Control Group and Related Issues in Clinical Trials* ; ICH E11 *Clinical Investigation of Medicinal Products in the Pediatric Population* ; *Guidelines for Clinical Evaluation by Therapeutic Category* (E12) ; *Clinical Evaluation* (E14) ; *Pharmacogenomics* (E15, E16).

Dans les recommandations ICH-E3, E8 et E9, la notion de critères d'évaluation est abordée de façon globale sans être précisée au cas par cas. Les textes recommandent seulement de définir précisément le critère utilisé.

Agences réglementaires pour les essais cliniques

La *Food and Drug Administration*

L'agence réglementaire américaine *Food and Drug Administration* (FDA) a publié un document de référence en 2007, intitulé *« Guidance for industry : Clinical trial endpoints for the approval of cancer Drugs and Biologics »*. Ce document est le premier d'une série de recommandations sur les critères en cancérologie dans le but d'un dépôt d'autorisation de mise sur le marché (AMM) auprès de la FDA. Il décrit les critères principaux à utiliser dans quelle situation et précise les avantages et inconvénients *(tableau II)*.

Tableau II. Exemples de critères de survie utilisés dans les essais cliniques, selon la FDA [9].			
Critères	**Aspect réglementaire**	**Avantages**	**Inconvénients**
Survie globale	Bénéfice clinique pour les autorités réglementaires	Universellement accepté comme mesure directe du bénéfice Facile à mesurer Mesure précise	Peut nécessiter une large étude (nombre de sujets importants) Peut être affecté par le changement de traitement ou les traitements séquentiels Inclut les décès non liés au cancer
Survie sans maladie	Critère de substitution pour les autorisations accélérées ou autorisations réglementaires habituelles	Nombre de sujets plus faible et suivi plus court comparés aux études de survie globale	Critère de substitution de la survie globale non validé formellement pour toutes les localisations de cancer Mesure non précise et sujette au biais d'évaluation, notamment dans les études en ouvert Les définitions varient selon les études
Survie sans progression (inclut tous les décès) Ou Temps jusqu'à progression (les décès avant la progression sont censurés)	Critère de substitution pour les autorisations accélérées ou autorisations réglementaires habituelles	Nombre de sujets plus faible et suivi plus court comparés aux études de survie globale Comprend la mesure de la maladie stable Non affecté par le changement de traitement ou les traitements séquentiels Doit être basé sur une évaluation objective et quantitative	Critère de substitution de la survie globale non validé formellement pour toutes les localisations de cancer Mesure non précise et sujette au biais d'évaluation, notamment dans les études en ouvert Les définitions varient selon les études Fréquence des évaluations (radiologiques ou autres) Nécessite des évaluations à des temps similaires dans les bras de traitement

L'agence réglementaire pour les médicaments en Europe

De la même façon, l'agence européenne a participé à la production des recommandations internationales tripartites entre l'Union européenne, le Japon et les États-Unis, les « *International Conference on Harmonisation (ICH) of technical requirements for registration of pharmaceuticals for human use* » sur différents aspects notamment la mesure de l'efficacité (*Topic E1 to E16*, E pour *efficacy*). Ces recommandations sont relayées sur le site Internet de cette agence[2].

Les survies, en pratique de cancérologie

De façon schématique et non consensuelle, nous proposons un tableau *(tableau III)* pour l'utilisation des survies en fonction de situations cliniques en cancérologie, notamment pour les tumeurs solides. Ce tableau non exhaustif et simplifié permet à des « nouveaux venus » en cancérologie d'avoir un guide. Ce guide est et doit être discuté avec le porteur du projet en fonction des objectifs et des hypothèses de la recherche. Chaque critère peut être utilisé mais le choix d'un critère de substitution en tant que critère principal d'un essai clinique doit être argumenté en fonction de la validation formelle de ce critère. De fait, une recherche bibliographique comme une discussion avec un statisticien doit être réalisée au préalable.

Tableau III. Critères de survie utilisés en phase III selon les situations thérapeutiques pour les tumeurs solides.

Situation clinique	Critères de survie utilisés (français et anglais)
Néo-adjuvant	Survie globale ou OS Survie sans rechute ou RFS Survie sans maladie ou DFS
Adjuvant	Survie globale ou OS Survie sans maladie ou DFS
Avancé/métastatique	Survie globale ou OS Survie sans progression Temps jusqu'à progression Survie sans rechute ou RFS Survie sans rechute locale ou L-RFS Survie sans rechute métastatique ou M-RFS

OS : *Overall Survival* ; RFS : *Relapse free survival* ; DFS : *Disease Free survival* ; L-RFS : *Local Relapse free survival* ; M-RFS : *Metastasis Relapse free survival*.

2. http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm

À retenir pour les survies composées de plusieurs événements

- Idéalement, les événements non mortels doivent être en rapport avec l'histoire naturelle du cancer et avec l'issue d'une maladie.
- Chacun des composants d'un critère de survie doit être décrit, mais pas forcément testé (attention aux tests multiples). Des effets négatifs sur un ou plusieurs événements ne doivent pas être masqués par des effets fortement positifs sur les autres événements le constituant.
- Les patients ayant un premier événement non mortel ne doivent pas être exclus de l'étude. Il ne faut donc pas arrêter le suivi d'un patient après la survenue du premier événement d'intérêt.
- Penser à l'utilisation des risques compétitifs et lire le chapitre III.3 « Événements à risque compétitif », page 164.
- S'assurer de la validité du critère de survie utilisé en tant que critère principal de l'essai clinique et consulter les recommandations internationales.
- Savoir que les définitions peuvent varier d'un article à l'autre pour les survies composées de plusieurs événements. Il faut donc vérifier dans la partie « méthodes » des articles quels événements ont été pris en compte pour les estimations, il en est de même pour les événements censurés.

Références

1. Estève J, Benhamou E, Croasdale M, Raymond L. Relative survival and the estimation of net survival: Elements for further discussion. *Statistics in Medicine* 1990 ; 9 (5) : 529-38.

2. Conroy T, Desseigne F, Ychou M, *et al.* Randomized phase III trial comparing FOLFIRINOX (F: 5FU/leucovorin [LV] irinotecan [I], and oxaliplatin [O]) versus gemcitabine (G) as first-line treatment for metastatic pancreatic adenocarcinoma (MPA): Preplanned interim analysis results of the PRODIGE 4/ACCORD 11 trial. *Clin Oncol* 2010 ; 28 (suppl) : 15s (abstr 4010).

3. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, *et al.* Trastuzumab after Adjuvant Chemotherapy in HER2 Positive Breast Cancer. *N Engl J Med* 2005 ; 353 : 1659-72.

4. The Breast International Group (BIG) 1-98 Collaborative Group. A Comparison of Letrozole and Tamoxifen in Postmenopausal Women with Early Breast Cancer. *N Engl J Med* 2005 ; 353 ; 26.

5. Montori VM, Devereaux PJ, Adhikari NK, *et al.* Validity of composite end points in clinical trials. *BMJ* 2005 ; 330 : 594-6.

6. Yothers G. Toward progression-free survival as a primary end point in advanced colorectal cancer. *J Clin Oncol* 2007 ; 25 (33) : 5218-24.

7. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, *et al.* Survival end point reporting in randomized cancer clinical trials: A review of major journals. *J Clin Oncol* 2008 ; 26 : 3721-6.

8. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001 ; 69 : 89-95.

9. FDA Guidance Documents. Guidance for industry : Clinical trial endpoints for the approval of cancer drugs and biologics. FDA, 2007 : http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071590.pdf

10. Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 2009 ; 27 : 2874-80.

11. Bhattacharya S, Fyfe G, Gray RJ, Sargent DJ. Role of Sensitivity Analyses in Assessing Progression-Free Survival in Late-Stage Oncology Trials. *J Clin Oncol* 2009 ; 27 : 5958-64.

12. Panageas KS, Ben-Porat L, Dickler MN, Chapman PB, Schrag D. When you look matters: The effect of assessment schedule on progression-free survival. *J Natl Cancer Inst* 2007 ; 99 (6) : 428-32.

13. Altman DG ; Schulz KF Moher D, *et al.* ; for the CONSORT Group. The revised CONSORT statement for reporting randomized trials : Explanation and elaboration. *Ann Intern Med* 2001 ; 134 (8) : 663-94.

14. Moher D, Hopewell S, Schulz KF, *et al*. Consolidated Standards of Reporting Trials Group. CONSORT 2010 explanation and elaboration : Updated guidelines for reporting parallel group randomised trials. *J Clin Epidemiol* 2010 ; 63 (8) : e1-37.

15. European Medicines Agency. International Conference on Harmonisation (ICH) of technical requirements for registration of pharmaceuticals for human use": http://www.emea.europa.eu/htms/human/humanguidelines/efficacy.htm

Critères de qualité de vie relatifs à la santé

F. Bonnetain, E. Chamorey, F. Kwiatkowski

La qualité de vie est un concept multidimensionnel incluant au minimum les domaines physiques, psychiques et sociaux ainsi que les symptômes liés à la maladie et aux traitements [1-5]. C'est un concept très large défini par les experts de l'Organisation mondiale de la santé (OMS) comme étant « la perception qu'a un individu de sa place dans la vie, dans le contexte de la culture et du système de valeurs dans lequel il vit, en relation avec ses objectifs, ses attentes, ses normes et ses inquiétudes », qui peut être influencée de manière complexe par la santé physique du sujet, son état psychologique, son niveau d'indépendance, ses relations sociales et sa relation aux éléments essentiels de son environnement [6]. D'autres auteurs préfèrent restreindre ce concept multidimensionnel à l'impact exercé par la maladie et les traitements sur le bien-être d'un individu [4]. D'autres retiennent que la qualité de vie est la résultante de la confrontation des espérances de santé d'un individu et de l'expérience qu'il en a ou que la qualité de vie traduit la capacité d'un individu à accomplir une vie pleinement satisfaisante [3]. Kaplan et Bust ont proposé le terme de « qualité de vie relative à la santé » pour distinguer l'influence des effets de la santé sur la qualité de vie des autres facteurs qui ne sont pas directement rattachés à la santé [7]. Au final, la qualité de vie résume globalement les jugements qu'ont les individus de leur santé et de leur maladie.

Ce chapitre traite de la qualité de vie relative à la santé utilisée comme critère de jugement en recherche biomédicale. Il est constitué de quatre parties qui abordent les définitions, les propriétés psychométriques des questionnaires, la mise en place d'une étude et la méthodologie d'analyse. Pour faciliter la lecture, l'acronyme français de « qualité de vie » (QdV) a été utilisé pour désigner la qualité de vie relative à la santé.

Définitions

Typologie des instruments de mesure

État de santé *versus* préférences du patient

Il existe deux grands types de mesure de la QdV résultant de la différence de perspectives de deux disciplines [4] :

• **les modèles psychométriques** qui mesurent l'état de santé : ils ont pour vocation de comparer des groupes de patients dans le cadre d'essais cliniques et de mesurer un changement au cours du temps ;
• **les modèles de préférences économétriques** qui relèvent du concept d'utilité attribué à différents états de santé : ils sont utilisés pour mesurer la préférence entre la qualité et la quantité de la vie. Les plus connues sont les QALY's et le *Health utility index* [8].

Mesures objectives ou mesures subjectives

On peut considérer qu'il existe deux types de mesures de santé qui s'opposent : les mesures objectives (poids, taille, surface corporelle) et les mesures subjectives (QdV, échelle de dépression, etc.). Cette nature subjective dessert l'évaluation de la QdV car la mesure objective, quantifiable, est considérée comme plus valide, fiable et reproductible. Cependant, quand on examine précisément ces deux types de mesures, cette différence ne paraît pas aussi évidente. La subjectivité de la QdV est relative, elle est à replacer dans un faisceau de mesures cliniques utilisées en routine. D'une part, de nombreux critères biomédicaux classiquement considérés comme objectifs sont de nature subjective : échelles de *Performance Status*, Karnofsky, grades de toxicité, progression, etc. De plus, de nombreuses mesures cliniques dites objectives peuvent être empreintes d'un degré d'erreur élevé, présenter une mauvaise concordance inter-observateur (*i.e.* entre les experts cliniciens) ou avoir une faible valeur prédictive ou pronostique [9]. La QdV correspond à l'évaluation d'un état subjectif perçu par le patient lui-même ou par un évaluateur externe. Les outils de mesure correspondant permettent de passer d'une mesure qualitative subjective à une mesure quantitative *via* des algorithmes de calcul relevant des mathématiques et les procédures de validation de plus en plus précises et rigoureuses.

Auto-évaluation et hétéro-évaluation

Il existe des instruments hétéro-évaluatifs et des instruments auto-évaluatifs :
• **auto-évaluation :** ce mode d'administration du questionnaire par le patient lui-même doit être privilégié afin de capter toute la quintessence de l'évaluation de la QdV [5, 10]. Cependant, il existe de nombreuses situations pour lesquelles le patient ne peut pas répondre lui-même [11] bien qu'il soit pertinent de mesurer sa QdV ;
• **hétéro-évaluation :** la mesure par un professionnel ne peut être considérée comme un *« Gold Standard »* car, dans l'absolu, il a une moins bonne connaissance de la façon dont le patient perçoit sa QdV et sa santé [10]. Cependant, l'hétéro-évaluation peut apporter une source d'informations pour les patients ayant un cancer dont l'évolution de la maladie peut s'associer à une incapacité à répondre.

Index globaux et échelles multidimensionnelles

Parmi les outils qui servent à mesurer la QdV, on distingue les index globaux et ceux qui permettent d'évaluer un profil multidimensionnel de QdV :
• **échelle globale :** utiliser un index global offre l'avantage d'une analyse simple et concrète pour l'aide à la décision comme, par exemple, le choix d'une prise en charge ou d'une thérapeutique [12, 13] ;

• **échelles multidimensionnelles :** elles sont à privilégier quand l'objectif est d'apprécier différents effets du traitement sur la QdV.

Échelles génériques et spécifiques

1. Il existe deux grands types d'instruments de mesure de la QdV [9, 14, 15] :

• **les échelles génériques** conçues pour mesurer la QdV chez des individus présentant ou non une pathologie, c'est-à-dire quel que soit leur état de santé ;

• **les échelles spécifiques** se focalisent sur l'impact d'une pathologie et de ses traitements sur la QdV. Celles-ci sont souvent plus sensibles à un changement minime dû à un traitement [16]. Ces instruments peuvent être spécifiques d'une localisation d'un cancer, d'une population de patients, de certaines fonctionnalités, de certains symptômes.

Propriétés psychométriques requises d'un questionnaire de qualité de vie

La volonté de reconnaître la QdV comme critère de jugement en cancérologie et de la prendre en compte dans l'évaluation des politiques de santé et des thérapeutiques impose d'améliorer les fondements métrologiques, c'est-à-dire de s'assurer d'avoir des instruments « fiables » et valides pour la mesurer [17, 18]. La validation d'un instrument de mesure doit permettre de répondre à deux questions fondamentales : quelle est le concept mesuré par l'instrument et quel est la qualité de la mesure ?

Le modèle de mesure

Le modèle de mesure est la transformation mathématique qui permet de transformer les réponses aux différents items pour obtenir un score final correspondant à la mesure.

La validité

La validité permet de vérifier que l'instrument est approprié pour mesurer ce qu'il est censé mesurer [19]. Cette notion de validité recouvre en réalité différents concepts [20] : structurelle, intrinsèque, et extrinsèque :
• **la validité structurelle** permet de valider si les items de l'instrument convergent vers la mesure d'un seul et même paramètre sous plusieurs dimensions physique, psychique ou sociale ;
• **la validité intrinsèque :**
– la validité de présentation correspond à l'acceptabilité de l'instrument. Un examen détaillé de l'instrument, du questionnaire par des experts et patients permet d'apprécier subjectivement si celui-ci mesure ce qu'il est censé mesurer,
– la validité de contenu est un concept proche de la validité de présentation permettant de vérifier que celui-ci couvre les concepts, domaines majeurs qu'il est censé mesurer ;

• **la validité extrinsèque :**

– la validité convergente évalue la concordance avec un *Gold-Standard*, c'est-à-dire un instrument de référence déjà validé explorant le même concept,

– la validité divergente permet de tester l'association de l'instrument avec d'autres paramètres fiables et/ou objectifs en émettant des hypothèses de divergence.

La fiabilité

Deux grands types de fiabilité (ou consistance interne) peuvent être explorés :

• **la reproductibilité** qui se décompose en réalité en deux propriétés :

– *la répétabilité* est la capacité d'un instrument à produire des résultats identiques sous des conditions identiques. Le plan expérimental dit *test-retest* consiste à administrer aux patients le même questionnaire à deux temps différents mais dans un intervalle court,

– *la reproductibilité* est la capacité d'un instrument à produire des résultats similaires sous des conditions proches mais pas parfaitement identiques. Dans le cas d'une hétéro-évaluation, on peut estimer la variabilité des mesures obtenues par plusieurs évaluateurs sur un même patient : on parle de « fidélité inter-juge » ;

• **la consistance interne** est une notion évaluant l'homogénéité des items : elle est fonction du nombre d'items et de leur covariance au sein d'une dimension ou d'un score explorant un concept spécifique (par exemple, la fatigue ou la dimension sociale). L'erreur aléatoire induite par la sélection des items pour une dimension ou une échelle est ainsi explorée [17, 19]. L'alpha de Cronbach est le coefficient communément utilisé pour explorer la consistance interne [21].

La sensibilité et la pertinence

La **sensibilité au changement** est définie comme la capacité d'un instrument à mesurer un changement sans qu'il soit nécessairement cliniquement significatif. Elle est une propriété métrique de l'instrument. La sensibilité est une propriété nécessaire mais insuffisante [14, 22]. Cette sensibilité peut être altérée par des effets plafonds ou des effets planchers. La plupart des statistiques explorant la sensibilité au changement comparent l'amplitude du changement avec l'écart-type de la mesure initiale [23, 24].

La **pertinence** est la capacité d'un instrument à mesurer un changement ayant un sens clinique pour le patient. Il s'agit d'une propriété clinique de l'instrument qu'il convient de distinguer de la sensibilité au changement [14, 22]. Jaeschke préfère la définir comme la plus petite différence dans un score que le patient perçoit comme un changement [25]. L'étude d'Osoba utilise une approche dite *anchor-based* pour capter la signification subjective d'un changement auprès de 357 patients ayant un cancer mammaire ou bronchique [26]. Ils ont utilisé le QLQ-C30 en comparant la taille de l'effet *(effect size)* entre trois groupes définis selon un nombre de points à l'aide d'un questionnaire de signification subjective – plus petit changement ayant un sens clinique (5 à 10), changement modéré (10 à 20) et changement élevé (plus de 20) – et qui sont couramment utilisés pour comparer des scores de QdV ou faire des hypothèses.

La validation transculturelle

La validation transculturelle d'un questionnaire développé initialement dans une autre langue va au-delà d'une simple traduction. Il importe de s'assurer qu'un processus méthodologique a été mis en œuvre pour garantir une bonne adaptation des items au regard de la langue mais aussi de la culture [19, 20, 27, 28].

Mise en place des études de qualité de vie en cancérologie : quand et comment ?

Quel rationnel justifie la mesure de la qualité de vie dans un essai en cancérologie ?

Il est important d'identifier les typologies d'essais en cancérologie pour lesquelles la QdV doit être mesurée et sous quelles conditions. De manière consensuelle, la QdV doit être considérée comme un critère majeur dans les essais en cancérologie en plus de la survie globale, de la survie sans progression et du taux de réponses si la différence de survie attendue est petite et/ou si une différence de QdV modérée à importante est attendue [29]. Aujourd'hui, en cancérologie, les nouveaux traitements permettent rarement d'espérer des différences significatives de survie globale. Aussi, des critères de substitution intermédiaires sont de plus en plus préconisés afin de réduire la durée des études et/ou de limiter le nombre de patients à recruter comme, par exemple, la survie sans maladie [30]. Mais peu d'entre eux sont validés [31, 32], ce qui pourrait suffire à justifier la mesure de la QdV dès que l'on utilise un critère de substitution comme critère de jugement principal. La *Food and Drug Administration* (FDA) se positionne favorablement sur l'intérêt des mesures subjectives [33], plus particulièrement sur le ressenti des patients vis-à-vis de leur prise en charge *(patient-reported outcomes)*. La FDA considère que la QdV est un critère évaluant le bénéfice clinique.

Ci-après, sont énumérées les conditions pour lesquelles l'utilisation de la QdV dans un essai de phase III est classiquement justifiée [29, 34] : **la QdV est le critère de jugement principal de l'essai** :
- dans les essais d'équivalence ou de non-infériorité ;
- dans les essais de thérapie adjuvante ;
- dans les essais en situation palliative ;
- dans les essais de dose-intensité ;
- quand les nouveaux traitements ont un impact réduit sur la survie à long terme.

Protocole de recherche et études de qualité de vie

La mise en œuvre de la QdV dans un essai clinique en cancérologie doit s'associer à un certain nombre de prérequis concernant le protocole. Tous les principes de conduite d'un essai clinique sont applicables, mais il y a des spécificités à intégrer. L'objectif est de limiter les biais de mesure, d'interprétation et d'intégrer la nature multidimensionnelle de la qualité de vie [4, 29].

Le rationnel

Le rationnel justifie l'énergie et les moyens requis pour l'évaluation de la QdV dans un essai en argumentant l'intérêt de sa mesure. Les investigateurs ont l'habitude de collecter et d'utiliser des données biologiques ou radiologiques en routine, leur motivation concernant ces données est quasi naturelle [4]. Pour les données de QdV, une plus grande motivation est requise, le rationnel en est le vecteur. Celui-ci doit répondre à un certain nombre de questions simples. Comment les résultats de la QdV vont-ils être utilisés pour déterminer l'efficacité comparative des bras de traitement ? Comment ces résultats pourront-ils changer la prise en charge clinique des patients ? En outre, il facilite la rédaction d'objectifs ciblés des recherches [4].

Les objectifs des recherches

Ils sont obligatoires, précis et spécifiques, un objectif du type « comparer la qualité de vie entre le bras A et B » n'est pas suffisant, notamment du fait de la nature multidimensionnelle de la QdV, pour guider l'évaluation et l'analyse. Les objectifs doivent identifier :

– les dimensions de QdV pertinentes à analyser ;

– la population d'intérêt en intention de traiter ou per-protocole ;

– la temporalité, c'est-à-dire l'intervalle de temps pertinent pour évaluer la QdV ;

– la mention précisant s'il s'agit d'un essai à visée exploratoire ou confirmatoire ;

– les hypothèses à tester en précisant d'abord s'il s'agit d'une simple mesure (ou d'une évolution) de la QdV ou de démontrer une équivalence ou une supériorité en proposant des hypothèses quantifiables. Ce dernier point permet en outre de calculer le nombre de sujets nécessaires ou la puissance.

La sélection des patients

Dans l'idéal, les patients éligibles doivent être ceux ciblés et retenus pour les autres critères de jugement. Dans la pratique, certaines conditions cliniques d'ordre physique, cognitif et/ou des spécificités socioculturelles limitent la faisabilité d'une auto-évaluation exhaustive de la QdV. Il faut alors prévoir une exclusion la plus restreinte possible de certains patients afin de garantir la validité des résultats [4]. Si une forte proportion de patients est susceptible de ne pas pouvoir répondre, il faut planifier une stratégie spécifique pour évaluer la QdV en utilisant des instruments hétéro-évaluatifs et/ou en autorisant l'hétéro-évaluation dès que le patient n'est plus capable de répondre par lui-même. En outre, dans ce type de situation, il peut être très utile d'effectuer

le remplissage des deux formes de questionnaire simultanément dès le début pour tous les patients et ce, afin d'évaluer leur concordance dans le contexte de l'étude et de garantir l'absence de biais ensuite.

Le schéma d'étude

Le schéma de l'étude doit être précisé : il s'agit de définir le nombre d'évaluations et les temporalités de la mesure de la QdV. Dans les essais en cancérologie, c'est souvent un schéma longitudinal qui est retenu car il permet d'étudier de façon comparative les changements de QdV [29] dus à l'impact du traitement et à l'évolution de la maladie. L'*European Organization for Research and Treatment of Cancer* (EORTC) préconise au minimum trois temps de mesure : à l'inclusion, au moins un durant le traitement et à la fin du traitement [35, 36] :

• **à l'inclusion :** cette mesure permet de comparer la QdV selon le bras avant le traitement et éventuellement d'ajuster si des différences existent. Elle doit de préférence être réalisée avant la randomisation et l'initiation du traitement car sa connaissance peut influencer l'évaluation de la QdV [37]. De plus, la QdV à l'inclusion est un facteur pronostique de critères cliniques que sont la survie, la réponse aux traitements et les toxicités [38] ;

• **durant le traitement :** afin de limiter la logistique et d'améliorer l'observance *(compliance)*, ces mesures sont planifiées au même temps que les examens cliniques requis par l'essai. Ils doivent être choisis en tenant compte des toxicités et de l'efficacité du traitement. On distingue deux grands types de schéma d'évaluation [4] :

– les schémas reposant sur les événements *(Event-driven)* : quand l'objectif est de comparer la QdV des patients dans des conditions cliniques similaires, l'évaluation doit être planifiée au regard d'événements cliniques pertinents ou correspondre à des phases spécifiques de la prise en charge. Ce schéma est à préconiser quand les traitements comparés sont limités dans le temps et ont des schémas d'administration distincts en termes de temps et de durée,

– les schémas reposant sur des temporalités définies *(Time-driven)* consistent à planifier une mesure de la QdV à une fréquence définie temporellement. Cette méthode pose des problèmes de validité comparative des données de QdV, la trajectoire de prise en charge pouvant différer selon les patients. Elle est plus particulièrement adaptée quand les traitements sont délivrés selon le même schéma et quand l'essai porte sur une période de temps plus longue ;

• **la fin du traitement :** la QdV devra être mesurée régulièrement après le traitement si elle est le critère de jugement principal et/ou si l'on s'intéresse à l'impact à long terme du traitement sur la QdV ;

• **après un arrêt prématuré du traitement :** un patient peut arrêter un traitement pour cause de toxicité et/ou de progression de sa maladie. Si sa QdV n'est plus évaluée, cela pourrait engendrer alors un biais différentiel. Ainsi, un bras de traitement avec un nombre plus important de patients sortis d'étude pourrait apparaître comme artificiellement bénéfique. En effet, seuls ceux en bonne santé auraient leur QdV évaluée. Toutefois, la mesure pourra être stoppée si l'objectif est de comparer la QdV durant le traitement.

Le choix de l'instrument de qualité de vie

Le choix du questionnaire repose sur l'objectif de l'étude et sa population. En outre, il faut prendre en considération les propriétés psychométriques, les méthodes de passation pour choisir l'outil adéquat. Un certain nombre d'auteurs ont proposé des check-lists pour sélectionner le ou les instruments adaptés à la recherche. Voici un ensemble de questions devant aider à guider le choix du questionnaire [4, 39] :

- L'instrument mesure-t-il ce qu'il est censé mesurer ?
- L'information sera-t-elle pertinente pour les objectifs de l'étude ?
- L'instrument évalue-t-il des dimensions de la QdV importantes pour cette étude ?
- Faut-il un instrument générique et ou spécifique ?
- L'instrument permettra-t-il de discriminer les patients de l'étude selon leur état clinique et pourra-t-il détecter un changement ?
- L'instrument peut-il servir à prédire d'autres critères de jugement ?
- Les questions sont-elles appropriées pour le patient durant toute la période de l'étude ?
- Le format et le mode d'administration sont-ils appropriés aux patients et à l'étude ?
- L'instrument a-t-il été initialement validé auprès d'une population similaire ? Sinon, quelles sont les mesures prises pour valider cet instrument auprès de cette population ?

La conduite d'un essai incluant la qualité de vie comme critère de jugement

Après avoir précisé le schéma, il est nécessaire d'organiser opérationnellement l'étude de la QdV. Cette phase est primordiale pour améliorer la qualité des données et prévenir les données manquantes.

Le mode d'administration

Auto- ou hétéro-évaluation, papier ou informatique, en établissement de santé ou à domicile, courrier papier ou électronique ? Le choix du questionnaire en fonction des objectifs et des patients prédétermine les modes d'administration et les moyens logistiques requis. Ils ont une influence sur la mesure de la QdV, ce qui rend nécessaire d'harmoniser les modalités de passation entre les bras [4, 40]. Il est communément admis que les patients cherchent à se conformer aux attentes du corps médical sur leur état de santé. C'est pourquoi il est impératif de préciser dans le protocole si le patient est autorisé à recevoir l'aide d'un tiers pour remplir un questionnaire auto-évaluatif (infirmière, médecin ou proche). Cette alternative ne devrait être, dans l'idéal, autorisée que dans les cas avérés d'impossibilité. Il demeure préférable d'obtenir une évaluation

de la QdV avec l'aide d'un tiers plutôt que des données manquantes. Il faut par ailleurs préciser à quel moment doit être rempli le questionnaire ; celui-ci doit être le même pour tous les patients et, en général, avant l'examen clinique avec le médecin [41].

La prévention des données manquantes

Malgré l'existence de techniques statistiques permettant de traiter les données manquantes, il est préférable d'avoir une action préventive afin d'éviter l'absence de données à chacune des étapes de la mise en place de l'étude [42]. Les causes majeures de ces données manquantes concernent d'abord les professionnels impliqués dans l'étude. Les sources de données manquantes sont souvent de plusieurs ordres, pouvant être dues :
– au décès, à la progression de la maladie ou aux toxicités sont difficilement évitables ;
– aux modalités de l'étude ;
– aux conditions logistiques et administratives de l'étude ;
– aux caractéristiques cliniques et sociales des patients.
Un ensemble d'actions préventives a été proposé de façon à prévenir les données manquantes [4, 34, 36, 42].

Pour les professionnels

Une étude pilote peut être envisagée dans chaque centre afin d'évaluer le rythme de recrutement, les moyens logistiques et financiers requis.

• Une formation spécifique peut être envisagée pour expliquer le protocole et les consignes.

• Il est primordial d'identifier dans chaque centre une personne responsable de la distribution et du recueil des questionnaires de QdV.

• Les procédures et le planning du recueil des questionnaires de QdV doivent être écrits explicitement et mis à disposition de chaque personne du centre participant à l'étude.

• Des rapports réguliers sur le recrutement et l'observance doivent être réalisés puis communiqués auprès des investigateurs et au comité indépendant de l'étude.

• Un item renseignant les raisons de non-remplissage du questionnaire de QdV doit être adjoint au cahier d'observation. Il aidera à analyser les déterminants des données manquantes et à optimiser le recueil en cours d'étude. Ces données seront analysées au regard des taux de complétion calculés pour une recherche analytique.

Pour les patients

• Donner les instructions pour le remplissage du questionnaire.

• Expliquer clairement l'importance de la mesure de la QdV dans le cadre du protocole.

• Expliquer et décrire à quels moments il sera demandé au patient d'évaluer sa QdV.

• S'assurer et expliquer que les informations sur la QdV resteront anonymes.

• S'assurer que le questionnaire lui-même n'est pas trop long, que les questions sont acceptables et pertinentes pour le patient.

• Offrir un environnement propice au remplissage du questionnaire.

• Proposer de l'aide si nécessaire pour compléter le questionnaire.

Méthodologies pour l'analyse de la qualité de vie

Comme pour tous les critères de jugement, il faut initialement rédiger un plan d'analyse. Un autre enjeu est de produire des résultats qui puissent être interprétés et exploités par les cliniciens. La production de résultats comparables entre les études est un enjeu majeur pour le développement de l'utilisation de la qualité de vie en cancérologie [18].

Population analysée

L'analyse peut être réalisée en intention de traiter, per-protocole et/ou auprès de sous-groupes de patients. Plus les patients analysés diffèrent de la population randomisée, plus la probabilité d'être confronté à un biais de sélection est grande.

Modèles d'analyse longitudinale

Les deux modèles les plus couramment utilisés sont [4] : les modèles à mesures répétées et les modèles à variables continues.

Modèles à mesures répétées

Dans un modèle à mesures répétées, le temps est considéré comme une variable catégorielle. Ces modèles sont préconisés dans les études longitudinales avec un nombre limité de suivi (2 à 5), l'évaluation de la QdV pouvant être réalisée dans un intervalle de temps prédéfini. Il devient plus difficile à utiliser si la fréquence des évaluations est grande et si l'intervalle de temps entre les évaluations varie énormément entre les sujets [43, 44]. La façon la plus simple de conceptualiser ce modèle est la régression linéaire. Étant donné que les évaluations répétées de la QdV sont corrélées entre elles, il faut donc préciser la structure de covariance qui peut être non structurée *(unstructured)* ou structurée (du type autorégressive). En général, il s'agit d'un modèle paramétrique linéaire à effets mixtes correspondant à la prise en compte d'effets fixes et d'effets aléatoires [4, 45, 46]. Si on note Y_i le vecteur des scores de QdV pour le sujet *i*, le modèle est défini par :

$$Y_i = X_i\beta + Z_i\gamma_i + \varepsilon_i$$

où X_i est une matrice de variables explicatives qui peut inclure le temps et des variables dépendantes ou non du temps, et Z_i une sous-matrice de X_i, β est un vecteur d'effets fixes et γ_i est un vecteur d'effets aléatoires de distribution normale, de moyenne 0 et de matrice de covariance B. Le vecteur ε_i suit une distribution normale, de moyenne 0 et de matrice de covariance V_i.

Modélisation des courbes de croissance

Ce sont les modèles où le temps est conceptualisé comme une variable continue *(growth curve model)* [4]. Ils utilisent le plus fréquemment des fonctions polynomiales. Ils sont à utiliser de préférence lorsqu'une grande variabilité individuelle de la temporalité de l'évaluation est observée et quand les temps de mesure sont nombreux. Afin de limiter les problèmes d'interprétation, tout en déviant de la linéarité, les polynômes modélisés devront être cubiques ou quadratiques. Du fait de mesures répétées, les données de QdV sont corrélées. Pour modéliser la structure de covariance, il est requis d'utiliser un modèle mixte. L'effet fixe modélise la moyenne de la QdV, et l'effet aléatoire la variation individuelle autour de cette moyenne. Dans un modèle longitudinal, il y a de préférence deux effets aléatoires. Le deuxième effet correspond à la variabilité des changements au cours du temps.

Temps jusqu'à détérioration de score de qualité de vie

En utilisant une modélisation de type survie à l'aide de modèle de Kaplan Meier, il est possible de produire des courbes de survie pour décrire longitudinalement la QdV. Cependant, les définitions des événements doivent être évaluées et être de préférence reproduites entre les études pour autoriser les comparaisons [47]. La définition de l'événement est donc le prérequis.

Conclusion

La QdV apparaît comme un critère majeur pour évaluer les prises en charge des patients en cancérologie et s'assurer qu'un bénéfice *a priori* clinique se traduit bien en un avantage de QdV. Il est important de poursuivre les recherches dans ce domaine et d'optimiser son évaluation, son analyse et son interprétation afin qu'elle puisse être utilisée comme un critère de jugement principal. À l'instar des autres critères de jugement tels que la survie sans progression, les clés de cet avènement sont l'utilisation de la QdV en pratique quotidienne, la production de résultats robustes et compréhensibles par le clinicien.

Références

1. Wood-Dauphinee S. Assessing quality of life in clinical research: From where have we come and where are we going? *J Clin Epidemiol* 1999 ; 52 (4) : 355-63.

2. Health outcomes methodology. *Med Care* 2000 ; 38 (9 Suppl) : II7-13.

3. Carr AJ, Gibson B, Robinson PG. Measuring quality of life : Is quality of life determined by expectations or experience ? *BMJ* 2001 ; 322 (7296) : 1240-3.

4. Fairclough DL. *Design and analysis of quality of life studies in clinical trials.* Boca Raton: Chapman and Hall, 2002, 328 pages.

5. Osoba D. Lessons learned from measuring health-related quality of life in oncology. *J Clin Oncol* 1994 ; 12 (3) : 608-16.

6. WHOQOL Division of mental health World Health Organisation. Study protocol: Internal document (NNH/TSS/93.9, 1993).

7. Kaplan RM, Bush JW. Health-related quality of life measurement for evaluation research and policy analysis. *Health Psychology* 1982 ; 1 : 61-80.

8. Torrance GW, Feeny DH, Furlong WJ, *et al.* Multiattribute utility function for a comprehensive health status classification system. Health Utilities Index Mark 2. *Med Care* 1996 ; 34 (7) : 702-22.

9. Wiklund I, Karlberg J. Evaluation of quality of life in clinical trials. Selecting quality-of-life measures. *Control Clin Trials* 1991 ; 12 (4 Suppl) : 204S-216S.

10. Gill TM, Feinstein AR. A critical appraisal of the quality of quality-of-life measurements. *JAMA* 1994 ; 272 (8) : 619-26.

11. Addington-Hall J, Kalra L. Who should measure quality of life? *BMJ* 2001 ; 322 (7299) : 1417-20.

12. Sloan JA, Loprinzi CL, Kuross SA, *et al.* Randomized comparison of four tools measuring overall quality of life in patients with advanced cancer. *J Clin Oncol* 1998 ; 16 (11) : 3662-73.

13. Sloan JA, Aaronson N, Cappelleri JC, *et al.* Assessing the clinical significance of single items relative to summated scores. *Mayo Clin Proc* 2002 ; 77 (5) : 479-87.

14. Guyatt GH, Deyo RA, Charlson M, *et al.* Responsiveness and validity in health status measurement : A clarification. *J Clin Epidemiol* 1989 ; 42 (5) : 403-8.

15. Guyatt GH, Kirshner B, Jaeschke R. A methodologic framework for health status measures: Clarity or oversimplification ? *J Clin Epidemiol* 1992 ; 45 (12) : 1353-5.

16. Wiebe S, Guyatt G, Weaver B, *et al.* Comparative responsiveness of generic and specific quality-of-life instruments. *J Clin Epidemiol* 2003 ; 56 (1) : 52-60.

17. Hamon A, Mesbah M. Internal statistical validation of a quality of life questionnaire. *Rev Epidemiol Sante Publique* 1999 ; 47 (6) : 571-83.

18. Lipscomb J, Donaldson MS, Arora NK, *et al.* Cancer outcomes research. *J Natl Cancer Inst Monogr* 2004 ; (33) : 178-97.

19. Streiner DL, Norman GR. *Health measurement scales: A practical guide to their development and use, 2nd ed.* New York : Oxford Medical, 1995.

20. Hays RD, Anderson R, Revicki D. Psychometric considerations in evaluating health-related quality of life measures. *Qual Life Res* 1993 ; 2 (6) : 441-9.

21. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951 ; 16 : 297-334.

22. Terwee CB, Dekker FW, Wiersinga WM, *et al.* On assessing responsiveness of health-related quality of life instruments: Guidelines for instrument evaluation. *Qual Life Res* 2003 ; 12 (4) : 349-62.

23. Deyo RA, Diehr P, Patrick DL. Reproductibility and responsiveness of health status measures statistics and strategies for evaluation. *Control Clin Trials* 1991 ; 12 (4 Suppl) : 142-58.

24. Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health status. *Med Care* 1989 ; 27 (3 Suppl) : S178-89.

25. Jaeschke R, Singer J, Guyatt GH. A comparison of seven-point and visual analogue scales. Data from a randomized trial. *Control Clin Trials* 1990 ; 11 (1) : 43-51.

26. Osoba D, Rodrigues G, Myles J, *et al*. Interpreting the significance of changes in health-related quality-of-life scores. *J Clin Oncol* 1998 ; 16 (1) : 139-44.

27. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: Literature review and proposed guidelines. *J Clin Epidemiol* 1993 ; 46 (12) : 1417-32.

28. Bowden A, Fox-Rushby JA. A systematic and critical review of the process of translation and adaptation of generic health-related quality of life measures in Africa, Asia, Eastern Europe, the Middle East, South America. *Soc Sci Med* 2003 ; 57 (7) : 1289-306.

29. Gotay CC, Korn EL, McCabe MS, *et al*. Quality-of-life assessment in cancer treatment protocols: Research issues in protocol development. *J Natl Cancer Inst* 1992 ; 84 (8) : 575-9.

30. Sargent DJ, Wieand HS, Haller DG, *et al*. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. J *Clin Oncol* 2005 ; 23 : 8664-70.

31. Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 2009 ; 27 : 2874-80.

32. Bonnetain F. Health related quality of life and endpoints in oncology. *Cancer Radiother* 2010 ; 14 (6-7) : 515-8.

33. *Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics*. U.S. Department of Health and Human Services Food and Drug Administration, 2007.

34. Moinpour CM, Feigl P, Metch B, *et al*. Quality of life end points in cancer clinical trials: Review and recommendations. *J Natl Cancer Inst* 1989 ; 81 (7) : 485-95.

35. Fayers PM, Hopwood P, Harvey A, *et al*. Quality of life assessment in clinical trials-guidelines and a checklist for protocol writers: The U.K. Medical Research Council experience. MRC Cancer Trials Office. *Eur J Cancer* 1997 ; 33 (1) : 20-8.

36. Young T, de Haes H, Curran D. *Guidelines for assessing quality of life in EORTC clinical trials*. Brussels: EORTC, 2004.

37. Brooks MM, Jenkins LS, Schron EB, *et al*. Quality of life as baseline is assessment after randomization is valid? *Med Care* 1998 ; 36 (10) : 1515-9.

38. Quinten C, Coens C, Mauer M, *et al*. Baseline quality of life as a prognostic indicator of survival: A meta-analysis of individual patient data from EORTC clinical trials. *Lancet Oncol* 2009 ; 10 : 865-71.

39. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality-of-life instruments: Attributes and review criteria. *Qual Life Res* 2002 ; 11 (3) : 193-205.

40. Bernhard J, Gusset H, Hurny C. Practical issues in quality of life assessment in multicentre trials conducted by the Swiss Group for Clinical Cancer Research. *Stat Med* 1998 ; 17 (5-7) : 633-9.

41. Osoba D, Zee B. Completion rates in health-related quality-of-life assessment : Approach of the National Cancer Institute of Canada Clinical Trials Group. *Stat Med* 1998 ; 17 (5-7) : 603-12.

42. Bernhard J, Cella DF, Coates AS, *et al*. Missing quality of life data in cancer clinical trials: Serious problems and challenges. *Stat Med* 1998 ; 17 (5-7) : 517-32.

43. Dabakuyo TS, Fraisse J, Causeret S, *et al*. A multicenter cohort study to compare quality of life in breast cancer patients according to sentinel lymph node biopsy or axillary lymph node dissection. *Ann Oncol* 2009 ; 20 (8) : 1352-61.

44. Bonnetain F, Bouche O, Michel P, *et al*. A comparative longitudinal quality of life study using the Spitzer quality of life index in a randomized multicenter phase III trial (FFCD 9102): Chemoradiation followed by surgery compared with chemoradiation alone in locally advanced squamous resectable thoracic esophageal cancer. *Ann Oncol*. 2006 ; 17 (5) : 827-34.

45. Cnaan A, Laird NM, Slasor P. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Stat Med* 1997 ; 16 (20) : 2349-80.

46. Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982 ; 38 (4) : 963-74.

47. Bonnetain F, Dahan L, Maillard L, *et al*. Time to definitive quality of life score deterioration as a means of longitudinal analysis for treatment trials in patients with metastatic pancreatic adenocarcinoma. *Eur J Cancer* 2010 ; 46 (15) : 2753-62.

Critères de substitution

X. Paoletti, F. Bonnetain

En cancérologie, le bénéfice thérapeutique de nouveaux traitements ou stratégies thérapeutiques se mesure, dans l'immense majorité des cas, sur le prolongement de la survie du patient. Or ce critère d'évaluation requiert, en situation de bon pronostic, un long suivi et de très nombreux patients pour obtenir une puissance statistique satisfaisante. C'est particulièrement le cas des traitements administrés après une chirurgie à visée curative où la chimiothérapie adjuvante doit permettre de prévenir les rechutes. Outre un coût élevé lié à une surveillance régulière à long terme, l'allongement de la période de suivi a pour conséquence de retarder les conclusions de l'essai et la mise à disposition à l'ensemble des patients d'une thérapie potentiellement efficace.

Il est donc intéressant de pouvoir remplacer un critère final par un critère « intermédiaire » qui pourrait être mesuré plus précocement. Le lecteur trouvera différentes définitions du critère de substitution dans le *tableau I*. De manière plus générale, tout critère pouvant être mesuré plus rapidement, avec des techniques moins invasives et/ou à un coût moindre, est un bon candidat *(tableau I)*.

On qualifie également ces critères de « substitutifs » si, au préalable, ils démontrent des qualités de prédiction de l'effet du traitement suffisantes. Or si, dans la littérature, de nombreux critères autres que la survie sont utilisés pour mesurer le bénéfice d'une approche thérapeutique, ils n'ont pas toujours été validés correctement, conduisant les agences réglementaires à une méfiance importante vis-à-vis de ceux-ci. Ainsi, une suite de décisions erronées prises par la *Food and Drug Administration* (FDA) sur la base de critères de substitution a fortement contribué à les discréditer. Nous soulignons donc l'importance d'une validation rigoureuse par des méthodes statistiques *ad hoc* de ces critères de substitution avant leur utilisation comme critère principal.

Si une certaine « corrélation » entre le critère final et son substitutif est nécessaire, ce n'est pas une condition suffisante comme souligné par Fleming et DeMets [1] : les exemples sont nombreux où un traitement efficace sur un critère intermédiaire s'est révélé inefficace, voire délétère sur le critère clinique final. La *figure 1* fournit l'illustration qu'un critère peut être fortement pronostique du critère final sans être un critère de substitution et *vice versa* ! La mesure de cette association et la validité d'un critère substitutif requièrent des méthodes statistiques particulières et souvent délicates à mettre en œuvre.

Tableau I. Différentes définitions d'un critère de substitution dans la littérature.
• Un critère de substitution est une variable S pour laquelle tester l'hypothèse nulle « le traitement n'a pas d'effet sur S » équivaut à tester l'hypothèse nulle « le traitement n'a pas d'effet sur le critère d'efficacité clinique de référence » [2]. • Un critère de substitution est une mesure de laboratoire ou un signe physique utilisé dans les essais cliniques à la place d'un critère cliniquement pertinent pour le patient [...] et qui est supposé prédire l'effet du traitement. Un effet sur le critère de substitution n'a pas en lui-même de valeur pour le patient [3]. • Un critère de substitution dans un essai clinique est une mesure de laboratoire ou un signe physique utilisé à la place d'un critère cliniquement pertinent pour le patient [...]. Les changements provoqués par le traitement sur le critère de substitution sont supposés refléter les changements sur le critère clinique [4]. • Un critère de substitution est un biomarqueur qui peut se substituer au critère de jugement clinique. Selon des arguments épidémiologiques, thérapeutiques, physiopathologiques ou d'autres arguments scientifiques, un critère de substitution doit permettre de prédire le bénéfice (ou risque) clinique (ou l'absence de bénéfice ou de risque) [5-7]. • Un critère de substitution est un critère intermédiaire mesurable de façon fiable et reproductible (imagerie, biologie, etc.), et qui permet de prédire l'effet du traitement sur le critère clinique.

Nous nous restreindrons ici à la situation des essais randomisés de phase III construits pour évaluer le bénéfice clinique d'une intervention donnée. De fait, nous n'aborderons pas les autres étapes du développement clinique comme les études de phase I ou de phase II, où sont utilisés des critères différents, potentiellement substitutifs. Par exemple, en phase I, la dose optimale repose sur la mesure de la toxicité car on suppose une forte corrélation entre l'activité antitumorale et la toxicité, du moins pour les traitements cytotoxiques. D'autre part, les critères de substitution ne sont pas l'apanage de la cancérologie, et le contenu de ce chapitre peut aisément se transposer à d'autres disciplines médicales.

Deux cadres différents ont été utilisés pour développer des concepts et des méthodes de validation des critères de substitution. Historiquement, les premiers travaux ont cherché à exploiter l'information apportée par un essai unique. Puis l'approche dite méta-analytique a offert un cadre plus large avec de nouvelles notions permettant de mieux appréhender la validation des critères de substitution. Pour suivre cette chronologie, nous présenterons en première partie les méthodes pour les essais uniques, en particulier trois concepts fondamentaux qui sont le critère de Prentice, la proportion expliquée de Freedman ainsi que l'effet relatif et l'association ajustée. Nous étudierons ensuite l'approche méta-analytique telle que présentée par Buyse *et al.* [8, 9] ainsi que Burzykowski *et al.* [10]. Dans une dernière section, nous aborderons la délicate question de la planification d'un essai utilisant un critère de substitution.

La *figure 1* présente deux situations qui montrent une relation entre l'effet du traitement sur le critère final (T) et l'effet du traitement sur le critère de substitution (S). Chaque point représente la mesure de l'effet du traitement 1 ou 2.

Dans la situation a, le critère S et le critère T sont fortement corrélés (points alignés le long de l'axe en pointillés et de la même manière dans les deux groupes de traitement). Le critère S

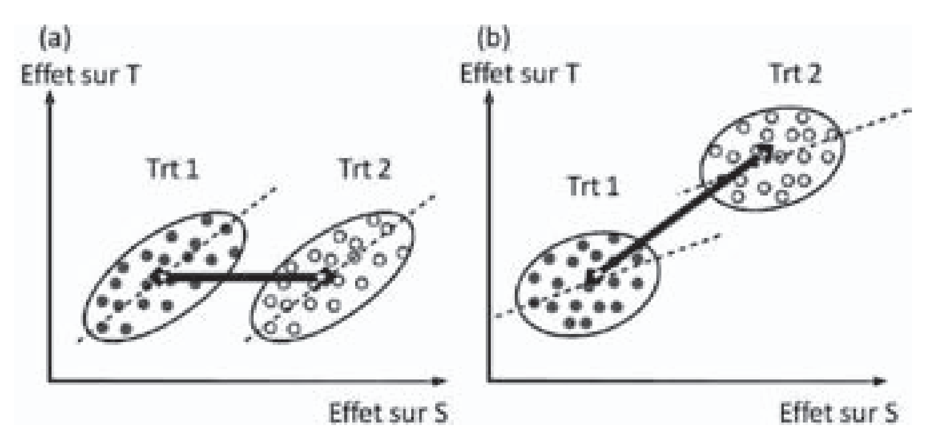


Figure 1. Facteur pronostique et critère de substitution.
Trt : traitement ; T indique le critère final, S le critère de substitution.

peut alors être considéré comme pronostique du critère T. En revanche, le traitement 2 augmente l'effet sur le critère S en moyenne, mais n'a aucun impact sur la moyenne de l'effet de T. Un effet sur le critère S ne s'accompagne pas d'un effet sur le critère T. On considère que le critère S n'est pas un critère de substitution du critère clinique T.

Dans la situation b, le critère S et le critère T sont moins bien corrélés : le nuage de points est moins resserré autour de la droite avec une pente plus faible. Cependant, le critère S peut être considéré comme un critère de substitution du critère T dans la mesure où l'augmentation de l'effet sur le critère S provoqué par le traitement 2 s'accompagne d'une augmentation de l'effet sur le critère T.

Validation sur un essai unique

Critère de Prentice

En 1989, Prentice *et al.* [2] ont proposé une définition des critères de substitution qui a fortement contribué à structurer la réflexion. Cette définition a ensuite été déclinée en critères opérationnels pour aider à la validation de critères de substitution.

Selon la définition de Prentice [2], un critère de substitution est une variable de réponse pour laquelle un test de l'hypothèse nulle de l'absence de relation au traitement est également un test valide de l'hypothèse nulle correspondante pour le critère final (ou vrai critère). Si on suppose une variable traitement binaire Z (par exemple, Z = 1 si traitement expérimental et Z = 0 pour le contrôle) et qu'on dénote S la variable pour le critère de substitution et T la variable pour le critère final, cette définition peut s'écrire mathématiquement :

$$f(Z|S) = f(S) \Leftrightarrow f(T|S) = f(T) \quad (1)$$

où f(Z) dénote la distribution de probabilité de Z et f(Z | S) la distribution conditionnelle. Cette définition est de peu d'utilité pratique puisqu'elle requiert de nombreuses expériences qui pourraient ne pas être toutes vérifiées du fait des fluctuations d'échantillonnage ou d'un manque de puissance de certaines d'entre elles.

Quatre critères opérationnels ont donc été proposés pour vérifier que le triplet (Z, T, S) satisfait cette définition :

$$\begin{aligned} &1) \quad f(S|Z) \neq f(S) \\ &2) \quad f(T|Z) \neq f(T) \\ &3) \quad f(T|S) \neq f(T) \\ &4) \quad f(T|S,Z) = f(T|S) \end{aligned} \qquad (2)$$

1) effet du traitement statistiquement significatif sur le critère de substitution ;

2) effet du traitement statistiquement significatif sur le critère final ;

3) effet pronostique statistiquement significatif du critère de substitution ;

4) après ajustement sur le critère de substitution, l'effet du traitement sur le critère final n'est plus statistiquement significatif : le critère de substitution capture complètement l'effet du traitement.

Par exemple, pour un critère de substitution binaire S, qui prend deux valeurs, 0 et 1, la condition 2 équivaut à $\beta \neq 0$ dans le modèle semi-paramétrique de Cox $\lambda(t,Z) = \lambda_0 \exp(\beta Z)$ et la condition 4 équivaut à $\beta_S = 0$ dans le modèle $\lambda(t,Z) = \lambda_0 \exp(\beta_S Z + \gamma_Z S)$.

Quelques remarques sur ces critères opérationnels : un critère de substitution ne peut donc être validé que dans le cas où il existe un effet significatif du traitement sur T et Z, ce qui n'est pas le cas dans tous les essais. D'autre part, le 3e critère opérationnel indique simplement qu'il existe une association entre T et S, ce qui devrait évidemment être rempli par tout candidat. La dernière condition est celle qui capture le plus la notion de substitution : une fois pris en compte le critère de substitution, l'effet du traitement n'apporte plus aucune information sur le critère principal.

Les critiques de ces critères sont de deux ordres. La première est qu'il repose sur des tests d'hypothèse qui sont donc dépendants des erreurs de type I et II. Des conclusions erronées peuvent être tirées du fait uniquement d'un manque de puissance. La seconde, plus fondamentale, concerne le 4e critère opérationnel : Freedman *et al.* [11] rappellent que sa vérification revient à vouloir « prouver » qu'une hypothèse nulle est vraie. Enfin, une autre restriction à l'application d'un tel critère opérationnel est que l'effet du critère de substitution doit capturer totalement l'effet sur le critère final. Cela conduira à changer de perspective pour les travaux ultérieurs en se plaçant dans le cadre de l'estimation plutôt que dans celui des tests de comparaison.

En dépit de ces critiques, ce concept est attractif, notamment lorsqu'il existe un mécanisme biologique connu pour passer du critère intermédiaire au critère final.

Proportion expliquée de Freedman

Suite aux critiques précédentes, Freedman *et al.* ont proposé de calculer la proportion d'effet du traitement expliquée (PE) par le critère de substitution [11]. La PE est alors le ratio :

$$PE = 1 - \frac{\beta_s}{\beta} \tag{3}$$

où β est l'estimation de l'effet du traitement sur le critère final (par un modèle de régression par exemple) et β_s l'estimation de l'effet du traitement sur le critère final conditionnellement (ou après ajustement) sur l'effet du critère de substitution. Ainsi, le critère 4 de (2) correspond à PE = 1. L'intérêt de ce concept est, entre autres, de déplacer le problème de la validation du critère de substitution vers le champ de l'estimation et donc de la quantification. En quittant le cadre strict du test qui doit trancher entre « critère invalide *versus* critère de substitution parfait », il est possible d'apporter des réponses plus utilisables en pratique.

Toutefois, l'utilisation de la proportion expliquée pose un certain nombre de difficultés identifiées par les auteurs eux-mêmes. Tout d'abord, l'intervalle de confiance sera probablement assez large et va dépendre de la taille de l'effet sur le critère clinique final.

En outre, lorsque l'effet du traitement est faible, ce qui n'est pas rare en cancérologie, β approche de zéro, rendant le ratio très instable. De même, en cas d'interaction entre Z et T (par exemple, si le traitement a un effet différentiel sur le critère de substitution et sur le critère final), PE n'a plus une interprétation univoque puisqu'elle représente alors la proportion expliquée par le critère de substitution et par l'effet du traitement. Enfin, la dernière limite, et la plus fondamentale, est que la définition de PE ne correspond pas forcément à une proportion. Pour certains exemples, PE excède 100 % ou est négative comme cela a été décrit par Buyse et Molenberghs [9] dans une étude de validation du temps avant récidive comme critère de substitution de la survie globale dans le cancer du côlon résecable. Methy *et al.* [12] ont trouvé des résultats également aberrants lorsqu'ils ont étudié si le contrôle local pouvait être utilisé comme critère de substitution à la survie globale.

Ces résultats ont conduit Buyse et Molenberghs [9] à proposer deux concepts : l'effet relatif et l'association ajustée.

Effet relatif et association ajustée

Buyse et Molenberghs [9] ont considéré le ratio de l'effet du traitement sur le critère final et sur le critère de substitution (RE). Si on note α le paramètre associé à l'effet sur le critère final,

$$RE = \frac{\beta}{\alpha} \tag{4}$$

De même, en exprimant l'association entre les deux critères ajustée sur le traitement, ils ont mis en évidence que PE est composée de trois sources d'information :
• l'association ajustée qui est donc une mesure *au niveau individuel* : pour un patient, la valeur du critère intermédiaire est corrélée à celle du critère final ;
• le RE qui décrit la relation de l'effet du traitement sur les deux critères au niveau *de l'essai* ;
• le rapport des variances qui est un paramètre de nuisance mais qui peut servir de mesure de validation.
La mesure RE permet donc de prédire l'effet du traitement sur le critère final une fois observé l'effet sur le critère intermédiaire. La précision de cette prédiction dépendra de la précision de cette estimation.
En dépit de l'intérêt de pouvoir distinguer et quantifier les différentes composantes des relations dans le triplet (Z, S, T), la mesure du RE pose quelques problèmes. Le premier, qui est commun avec PE, est la faible précision de l'intervalle de confiance avec des essais de taille standard. Le second, plus limitatif, est que pour prédire la relation entre les deux critères pour un nouvel essai, ce qui est l'objectif *in fine*, il faut faire l'hypothèse d'un effet multiplicatif du traitement sur le critère de substitution et sur le critère final. Or cela est invérifiable sur un seul essai. L'approche méta-analytique apparaît alors naturelle.

Approche méta-analytique

Montrer qu'il existe une relation forte entre deux critères d'évaluation au niveau individuel, voire biologique, est essentiel et constitue un élément convaincant, notamment pour l'investigateur. Toutefois, lorsqu'il s'agit de modifier la prise en charge des patients sur la base d'un critère de substitution, les agences réglementaires ainsi que les statisticiens et les investigateurs souhaitent également savoir si la relation se maintient au niveau de l'essai. En d'autres termes, les conclusions d'un essai obtenues avec un critère intermédiaire permettent-elles de prédire les conclusions avec le critère clinique final ? Un essai unique ne peut apporter une réponse statistique à cette question. L'approche méta-analytique est alors nécessaire.

Il est possible de définir une autre unité que l'essai. Ainsi, ont été proposés ou utilisés :
• le centre, en excluant ceux ayant inclus uniquement des patients dans un même bras de traitement rendant impossible toute comparaison entre les traitements ;
• le centre ou l'essai, selon la taille de l'essai et l'information disponible sur les différents centres ;
• l'essai, en le décomposant en différentes comparaisons au bras de référence dans les essais multi-bras ;
• le pays, à l'intérieur de chaque essai ;
• l'investigateur.

Toutefois, comme illustré dans les paragraphes suivants, le choix de l'unité doit résulter d'un compromis entre le nombre des unités permettant d'estimer un effet du traitement et la précision de ces différentes estimations qui sont directement liées à la taille des unités.

Les notions de prédiction et d'erreur associée seront au centre de cette approche. Nous nous appuierons ici essentiellement sur les travaux de l'équipe de l'université de Hasselt par Marc

Buyse, Tomasz Burzykovski, Geert Molenberghs *et al.*, ainsi que sur ceux de Dan Sargent *et al.* Deux notions déjà évoquées dans la section précédente serviront de cadre : l'association au niveau individuel et au niveau de l'essai.

Régression linéaire pondérée

Green, Yothers et Sargent [13] décrivent un ensemble de méthodes de validation de critères de substitution et soulignent l'importance d'aborder les différentes facettes de la validation. Une méthode utilisée notamment par le groupe ACCENT pour la validation de la survie sans maladie (DFS pour *disease free survival*) en place de la survie globale dans le cancer du côlon est la régression d'un critère sur un autre sur un large ensemble d'essais. C'est une approche simple et intuitive tant pour les statisticiens que pour les cliniciens. Toutefois, elle n'est pas exempte de limites méthodologiques que nous aborderons à la fin de cette section. Nous nous restreindrons ici au cas des données de survie ; l'application aux variables binaires et continues est transposable.

Association au niveau individuel

L'idée est de mesurer le coefficient de corrélation entre les deux critères T et S à partir d'une régression linéaire pondérée par la taille de chacun des essais. Ainsi le taux de survie de S pour chaque bras de traitement de chaque essai est régressé sur le taux de survie de T pondéré par l'effectif de chaque bras. La variabilité de T expliquée par S est naturellement quantifiée par le coefficient R^2. Un coefficient de corrélation sur les rangs de Spearman peut également être calculé et donne *a priori* des valeurs proches. Un test des deux coefficients de régression permet de déterminer s'ils sont différents respectivement de 0 pour l'intercepte et de 1 pour la pente. En cas de pente statistiquement différente de 1, cela indique une atténuation (ou une augmentation) significative de l'effet entre le critère de substitution S et le critère clinique final T.

Du fait d'un suivi différent selon les essais inclus dans la méta-analyse, la régression peut conduire à des biais. Il est alors possible de mimer le déroulement de l'essai où le suivi des patients est variable selon leur date d'inclusion. Pour étudier la DFS à 3 ans par exemple, Sargent *et al.* [14] ont censuré les observations après un suivi médian de 3 ans.

Association au niveau de l'étude

Le même principe est appliqué pour l'association entre S et T sur l'effet du traitement. L'effet du traitement peut être modélisé par les *hazard ratios* (HR), comparant le bras expérimental au contrôle dans chaque essai, estimé par un modèle semi-paramétrique de Cox. Pour étudier l'association de l'étude, la relation entre les HR des deux critères est de nouveau estimée par un modèle linéaire pondéré par la taille respective de chacun des essais *(figure 2)*. Le coefficient de corrélation R^2 obtenu est maintenant spécifique de l'association au niveau de l'étude. Les mêmes tests que précédemment peuvent également être réalisés.

Une seconde mesure pour aider à juger de la similitude des conclusions obtenues avec chacun des deux critères est simplement les degrés de signification. Pour chaque essai, le test du log-rank conduit-il aux mêmes conclusions sur chacun des deux critères ? On obtient une mesure de concordance.

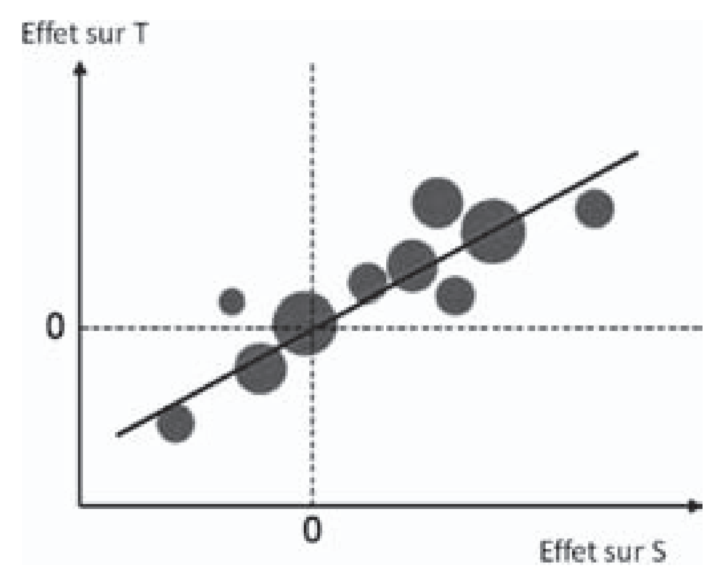


Figure 2. Approche méta-analytique : association entre les effets du traitement sur le critère clinique et le critère de substitution.
Les ronds les plus gros correspondent aux essais de plus grand taille.

Limites

Cette approche par modèle de régression univarié facilite une représentation graphique didactique et une interprétation intuitive, tout en permettant d'identifier de possibles observations extrêmes. Toutefois, comme souligné par Green *et al.* [13], cette approche réduit les données d'un bras de traitement à une observation pour chaque essai. Cela ne permet pas l'inclusion directe de covariables au niveau de l'essai et ne prend pas en compte le fait que chacun des HR est estimé avec erreur. La précision des prédictions obtenues avec la régression linéaire ainsi que le coefficient R^2 estimé peut donc être surestimée, en particulier si les études sont de taille modeste. En outre, le fait de calculer la survie à un temps donné semble limiter artificiellement l'information disponible puisque l'ensemble du suivi des différents patients n'est alors pas pris en compte. Certes, une telle approche est cohérente avec la planification d'une étude où généralement le critère principal est exprimé comme un taux à un temps donné, mais pour la validation d'un critère, ce ne semble pas être une nécessité.

Les résultats de concordance obtenus par l'analyse des degrés de signification ne sauraient être qu'auxiliaire puisque cette approche est totalement tributaire de la puissance des comparaisons avec chacun des critères. Ainsi, si le critère de substitution est plus sensible que le critère final, ce qui est une propriété recherchée, on peut mesurer un effet du traitement qui ne sera pas retrouvé sur le critère final du fait d'un manque de puissance. En outre, si les degrés de signification sont proches de 0,05, ils peuvent se trouver de part et d'autre de cette limite intangible conduisant à des conclusions opposées abusives.

Modélisation jointe des deux critères d'évaluation

Marc Buyse *et al.* [15] ont proposé une nouvelle approche pour étudier et valider des critères de substitution à partir de méta-analyses sur données individuelles. Pour chacun de ces deux niveaux, un modèle linéaire différent est utilisé où *i* dénote l'étude et *j* indice les patients :

$$\begin{aligned} S_{ij} &= \mu_{Si} + \alpha_i Z_{ij} + \varepsilon_{Sij} \\ T_{ij} &= \mu_{Ti} + \beta_i Z_{ij} + \varepsilon_{Tij} \end{aligned} \quad (5)$$

Leur association est capturée par la matrice de variance covariances des résidus.

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ & \sigma_{TT} \end{pmatrix}$$

Un modèle mixte à effets aléatoires réunissant les deux niveaux peut être utilisé de façon alternative au modèle à deux niveaux [16].

Association au niveau individuel

Pour étudier l'association au niveau individuel, Buyse et Molenbergs [9] ont suggéré de considérer l'association entre T et S après ajustement sur l'effet du traitement.

Du modèle précédent, ils déduisent que $(T_{ij}|Z_{ij}, S_{ij})$ suit une loi normale et calculent un coefficient de détermination au niveau individuel qui reflète l'association :

$$R^2_{indiv} = R^2_{\varepsilon_{Ti}|\varepsilon_{Si}} = \frac{\sigma^2_{ST}}{\sigma_{SS}\sigma_{TT}} \quad (6)$$

Cette mesure généralise la notion d'association ajustée mentionnée supra dans la section « Effet relatif et association ajustée » dans le cas d'une étude unique.

Association au niveau de l'étude

Au niveau de l'étude, la démarche est légèrement différente puisque cette fois, les auteurs s'intéressent à l'effet du traitement Z sur T, connaissant l'effet de Z sur S. À partir du modèle (5), il est possible de quantifier cette association par un autre coefficient de détermination, noté R^2_{essai}. Un critère de substitution parfait au niveau de l'essai donne une valeur égale à 1.

Il est donc possible, à partir de ces deux coefficients de détermination, non seulement d'évaluer la qualité du critère de substitution aux deux niveaux, mais également de prendre en compte les différentes sources de variabilité à partir de l'ensemble des données. En particulier, il n'est pas nécessaire de tronquer les observations de survie à un temps donné. Avant d'utiliser un critère de substitution en pratique, il faudrait montrer que ces deux coefficients sont élevés.

Limites

La principale restriction à l'utilisation de ce modèle est d'ordre numérique, car la convergence numérique peut être difficile à obtenir, si la variabilité intra- ou inter-étude n'est pas suffisamment importante. Un grand nombre d'études avec des effets du traitement différents sont préférables, même si, théoriquement, cette approche permet aussi de valider un critère de substitution lorsqu'il n'existe pas d'effet du traitement, du simple fait des fluctuations d'échantillonnage.

Planification d'un essai

Difficultés de la planification d'un essai utilisant un critère de substitution

Une fois un critère intermédiaire validé dans une localisation cancéreuse donnée, comment utiliser celui-ci ? La première réaction pourrait être de remplacer simplement dans le protocole un critère par un autre, en conservant le même objectif et les mêmes hypothèses à tester. On s'expose alors à deux risques :

• la différence cliniquement pertinente qu'on cherche à mettre en évidence peut être différente selon qu'on s'intéresse à un critère ou à un autre. De même, une absence d'effet du traitement sur un critère peut correspondre à un effet sur un autre critère. C'est ce qui était testé lorsqu'on confrontait les coefficients du modèle de régression univarié (pente et intercepte) à 1 et 0 respectivement. Une illustration de cette limite dans le cancer colorectal métastatique est présentée par Fleming *et al.* [17] ;

• la validation d'un critère pour une localisation, une situation et une classe thérapeutique peut ne pas se vérifier pour les traitements d'une nouvelle classe thérapeutique aux mécanismes d'action radicalement différents. Les récentes déconvenues du bevacizumab pour le traitement du cancer du sein, des ovaires ou du pancréas en situation avancée en sont des illustrations. Ainsi, si cette biothérapie a induit un bénéfice sur le temps avant progression, celui-ci ne s'est pas traduit par un bénéfice sur la survie globale [18]. Une interaction entre le traitement Z et l'association de S et T n'est jamais impossible *a priori*. En particulier, si le nouveau traitement modifie en profondeur la stratégie thérapeutique appliquée après la récidive qui, elle-même, a un impact sur la survie.

Flandre et O'Quigley ont proposé une approche en deux étapes pour étudier la relation entre S et T et utiliser cette relation chez les patients qui auraient atteint le critère intermédiaire pour augmenter la puissance de l'inférence sur le critère final [19]. Au cours de la première étape, les patients sont suivis pour le critère intermédiaire et le critère final afin de quantifier la force de l'association. Au cours de la seconde étape, un second groupe de patients est inclus et suivi uniquement pour le critère intermédiaire. Les deux critères sont finalement combinés, l'objectif étant de diminuer la durée globale de l'étude. Les deux étapes peuvent correspondre à deux essais effectués dans des contextes similaires. Les limites intrinsèques à l'approche sur un essai unique

s'appliquent. Toutefois cet article développe largement la notion de prédiction qui sera ensuite approfondie dans le concept d'effet seuil de substitution et dans des travaux de l'équipe de J. Taylor [20].

Effet seuil *(surrogate threshold effect)*

Pour faciliter la planification d'un essai, Burzykovski *et al.* ont proposé un nouvel outil qui mesure l'effet minimum à observer sur le critère intermédiaire pour prédire un effet significatif sur le critère final [21].

Utilisant le fait que les estimations du maximum de vraisemblance de la covariance sont asymptotiquement indépendantes des estimations des effets fixes dans les modèles linéaires mixtes, les auteurs fournissent un intervalle de confiance de la prédiction de l'effet sur le critère final conditionnellement à l'effet attendu sur le critère intermédiaire. Il est alors possible de spécifier un effet à observer sur S qui assure que l'intervalle de confiance de T n'inclura pas l'hypothèse nulle et donc qu'on obtiendra un effet statistiquement significatif sur le critère clinique final. C'est le *surrogate threshold effect* (STE). Il dépend de la relation entre S et T telle que présentée dans la section précédente, de l'erreur sur l'estimation de l'effet sur S, de l'erreur sur la prédiction de T à partir de S. Un STE large indique qu'il faut observer un effet important sur le critère intermédiaire pour être à même de tirer des conclusions robustes sur le critère final. La *figure 3* schématise cette relation pour déterminer le seuil Smin en se basant sur l'intervalle de confiance à 95 % de la droite de régression évalué au point zéro pour l'effet sur le critère T.

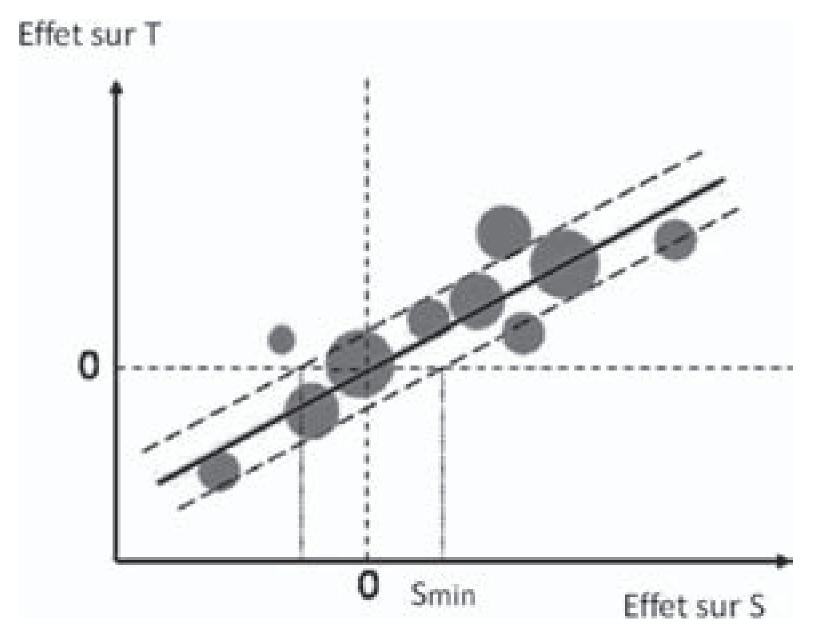


Figure 3. *Surrogate threshold effect* (STE) : effet minimum (Smin) à atteindre sur le critère de substitution S pour prédire un effet non nul sur le critère clinique T.

Une proposition de calcul de nombre de sujets nécessaires est en cours de développement par les auteurs. À titre d'exemple, il est recommandé pour le cancer du côlon métastatique d'observer une réduction du risque d'au moins 10 % sur la survie sans progression pour prédire une réduction du risque de 5 % sur la survie globale [10, 22].

Une des limites importantes à l'utilisation du STE est de connaître exactement la relation entre S et T ainsi que l'effet du traitement sur S, ce que seul le contexte de larges méta-analyses ainsi que de larges essais peut offrir. Une version approchée permet d'assouplir la première condition, mais outre le fait qu'elle repose sur des hypothèses d'une distribution normale de l'effet du traitement, la seconde condition est toujours maintenue. C'est sur la base de tels arguments que M. Gail [23] avait remis en cause l'usage des critères de substitution de façon générale.

Risques liés au changement de traitement sur la base des résultats intermédiaires

Du fait des limites sur la prédiction d'un effet par un autre, il est essentiel dans un essai de continuer à suivre les patients pour documenter le bénéfice sur le critère final (par exemple, la survie globale). En particulier, Fleming *et al.* [17] déconseillent fortement la mise sous traitement expérimental des patients du groupe contrôle sur la seule base des résultats sur le critère intermédiaire. En effet, lorsqu'une analyse intermédiaire montre un bénéfice du traitement à l'étude sur le temps avant récidive, il est assez commun que les patients qui progressent sous traitement contrôle se voient proposer le nouveau traitement. La mesure du bénéfice sur le critère final devient alors impossible. À moins que le traitement à l'étude ait fait la preuve de son bénéfice pour les malades en situation de rechute, une telle approche devrait être évitée.

Conclusions

Du fait de l'explosion de la durée et des coûts de la recherche clinique en cancérologie, la question de la recherche et de la validation des critères de substitution a été mise à nouveau sur le devant de la scène. Les critères de Prentice ont permis un formalisme constructif. Cette notion de validation a été de plus en plus exprimée comme un problème de prédiction d'un effet sur le critère principal à partir de l'effet sur le critère intermédiaire. Deux niveaux ont été distingués : le niveau individuel et le niveau de l'essai. L'approche méta-analytique semble être la seule à permettre cette double évaluation. De nombreuses mesures d'adéquation du modèle aux données ont été proposées pour quantifier la qualité de la substitution. Toutefois, elles montrent que, du fait d'erreurs de prédiction inhérentes aux modèles sur un nombre de données finies, l'usage des critères de substitution ne permet pas forcément des gains importants sur le nombre de sujets et la durée des études.

Rappelant que la validation d'un critère de substitution devait être réalisée dans une situation thérapeutique donnée et pour une classe pharmacologique donnée, Fleming a proposé [24] la classification suivante des critères de jugement :

• niveau 1 : critère de jugement mesurant un réel bénéfice clinique pour le patient ;
• niveau 2 : un critère de substitution validé selon l'approche méta-analytique ;
• niveau 3 : un critère de substitution non validé, mais qui a néanmoins raisonnablement des chances de prédire le bénéfice clinique souhaité, sur des considérations statistiques et cliniques.
• niveau 4 : une mesure biologique corrélée au critère clinique.

Le niveau 3 cristallise toutes les controverses, et il n'est pas clair que la distinction entre le niveau 3 et le niveau 4 soit pertinente. Cela pose la question du critère le plus efficace, c'est-à-dire qui permet de répondre à la question de façon convaincante mais dans des délais raisonnables. Ce n'est plus seulement une question statistique.

À retenir

• Un critère de substitution est un critère intermédiaire mesurable de façon fiable et reproductible (imagerie, biologie, etc.) et qui permet de prédire l'effet du traitement sur le critère clinique, dit critère final.
• Un critère n'est valide que pour une localisation, une situation et un traitement spécifique.
• Sa validation requiert la démonstration de ses bonnes corrélations avec le critère final tant au niveau individuel qu'au niveau de l'effet du traitement (niveau essai).
• La technique méta-analytique doit être privilégiée.
• Un critère intermédiaire non validé en tant que critère substitutif ne permet pas de déterminer le bénéfice clinique pour les patients.

Références

1. Fleming TR, DeMets DL. Surrogate end points in clinical trials: Are we being misled? *Ann Intern Med* 1996 ; 125 (7) : 605-13.

2. Prentice RL. Surrogate endpoints in clinical trials: Definition and operational criteria. *Stat Med* 1989 ; 8 (4) : 431-40.

3. Temple R. Are surrogate markers adequate to assess cardiovascular disease drugs? *JAMA* 1999 ; 282 (8) : 790-5.

4. Temple R. Policy developments in regulatory approval. *Stat Med* 2002 ; 21 (19) : 2939-48.

5. Biomarkers and surrogate endpoints: Preferred definitions and conceptual framework. *Clin Pharmacol Ther* 2001 ; 69 (3) : 89-95.

6. Lesko LJ, Atkinson AJ, Jr. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: Criteria, validation, strategies. *Annu Rev Pharmacol Toxicol* 2001 ; 41 : 347-66.

7. De Gruttola VG, Clax P, DeMets DL, *et al.* Considerations in the evaluation of surrogate endpoints in clinical trials. Summary of a National Institutes of Health workshop. *Control Clin Trials* 2001 ; 22 (5) : 485-502.

8. Aabo K, Adams M, Adnitt P, *et al.* Chemotherapy in advanced ovarian cancer: Four systematic meta-analyses of individual patient data from 37 randomized trials. Advanced Ovarian Cancer Trialists' Group. *Br J Cancer* 1998 ; 78 (11) : 1479-87.

9. Buyse M, Molenberghs G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* 1998 ; 54 (3) : 1014-29.

10. Burzykowski T, Buyse M, Yothers G, Sakamoto J, Sargent D. Exploring and validating surrogate endpoints in colorectal cancer. *Lifetime Data Anal* 2008 ; 14 (1) : 54-64.

11. Freedman LS, Graubard BI, Schatzkin A. Statistical validation of intermediate endpoints for chronic diseases. *Stat Med* 1992 ; 11 (2) : 167-78.

12. Methy N, Bedenne L, Conroy T, *et al*. Surrogate end points for overall survival and local control in neoadjuvant rectal cancer trials: Statistical evaluation based on the FFCD 9203 trial. *Ann Oncol* 2010 ; 21 (3) : 518-24.

13. Green E, Yothers G, Sargent DJ. Surrogate endpoint validation: Statistical elegance versus clinical relevance. *Stat Methods Med Res* 2008 ; 17 (5) : 477-86.

14. Sargent DJ, Wieand HS, Haller DG, *et al*. Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. *J Clin Oncol* 2005 ; 23 (34) : 8664-70.

15. Buyse M, Molenberghs G, Burzykowski T, Renard D, Geys H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000 ; 1 (1) : 49-67.

16. Molenberghs G, Buyse M, Geys H, Renard D, Burzykowski T, Alonso A. Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Control Clin Trials* 2002 ; 23 (6) : 607-25.

17. Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. *J Clin Oncol* 2009 ; 27 (17) : 2874-80.

18. Ocana A, Amir E, Vera F, Eisenhauer EA, Tannock IF. Addition of bevacizumab to chemotherapy for treatment of solid tumors : Similar results but different conclusions. *J Clin Oncol* 2011 ; 29 (3) : 254-6.

19. Flandre P, O'Quigley J. A two-stage procedure for survival studies with surrogate endpoints. *Biometrics* 1995 ; 51 (3) : 969-76.

20. Li Y, Taylor JM, Elliott MR. A bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* 2010 ; 66 (2) : 523-31.

21. Burzykowski T, Buyse M. Surrogate threshold effect: An alternative measure for meta-analytic surrogate endpoint validation. *Pharm Stat* 2006 ; 5 (3) : 173-86.

22. Buyse M, Burzykowski T, Carroll K, *et al*. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *J Clin Oncol* 2007 ; 25 : 5218-24.

23. Gail MH, Pfeiffer R, Van Houwelingen HC, Carroll RJ. On meta-analytic assessment of surrogate outcomes. *Biostatistics* 2000 ; 1 (3) : 231-46.

24. Fleming TR. Surrogate endpoints and FDA's accelerated approval process. *Health Aff* (Millwood) 2005 ; 24 (1) : 67-78.

Partie III
Analyses univariées

Données de survie

L. Campion, C. Bellera, A. Bajard

Pour bien appréhender et comprendre la thématique de ce chapitre, il est important de connaître les notions de probabilité, car les définitions des fonctions de survie, d'incidence cumulée et de probabilité instantanée prennent en compte le temps.

Un peu d'histoire

Les données de survie sont étudiées dès le XVIIe siècle par divers personnages fondateurs de cette théorie, issus d'horizons professionnels divers. John Graunt (1620-1674), marchand londonien et fondateur de la statistique démographique, conçoit les bases théoriques de l'élaboration des tables de mortalité grâce à l'étude des dénombrements de population et des *bulletins de mortalité de Londres* [1]. L'économiste William Petty (1623-1687) systématise et théorise les études démographiques sur les naissances, décès, nombre de personnes par famille, etc. L'astronome anglais Edmund Halley (1662-1742) établit en 1693 la première table de mortalité véritable en s'appuyant sur 5 ans d'état civil de la ville polonaise de Breslau (aujourd'hui Wroclaw). Suite à ces travaux d'experts, liés à des préoccupations pratiques, se développe au XIXe siècle l'utilisation financière large des données de survie par les actuaires des compagnies d'assurance.

Généralités

Les expressions « modèle de Cox », « test du log-rank », « estimation de Kaplan-Meier », « survie actuarielle » sont souvent rencontrées dans la littérature médicale internationale. Cette terminologie se rapporte à l'analyse statistique des données de survie ou, de façon plus générale, à des données liées au temps (délai de cicatrisation d'une plaie, délai jusqu'à une rechute de cancer, etc.). Dans de nombreuses affections et en particulier en cancérologie, la **durée de survie** globale est le critère principal d'évaluation thérapeutique. Analyser les données de survie, c'est s'intéresser à *l'apparition au cours du temps d'un événement spécifique, en l'occurrence du décès dans le cas spécifique de la survie globale*. En fait, tout événement de nature *binaire* (péjoratif ou non, médical ou non) peut être défini comme un critère d'intérêt : une rechute d'hyperthyroïdie, une panne mécanique, une guérison de leucémie. Mais la durée de survie a une particularité : *ce n'est pas*

une variable quantitative usuelle. Habituellement, on dispose pour chaque patient de la mesure du critère de jugement, et les calculs des moyennes et variances des deux échantillons sont possibles directement (*cf.* chapitre I.1 « Distributions statistiques », page 3). En revanche, la particularité des données de survie est l'existence de **données** « **incomplètes** » car *l'événement peut ne pas être observé pendant la période de suivi*. Dans ce cas, la durée de « vie » n'est pas connue au moment de l'analyse. La donnée de survie d'un patient qui n'est pas décédé (ou qui n'a pas eu l'événement considéré) au moment de l'analyse statistique est appelée « **donnée censurée** ».

Quelques exemples

Les deux exemples ci-dessous illustrent la représentation individuelle des données de survie avec des délais connus et censurés.

Illustration de la notion de censure

La *figure 1* représente la survie de 5 souris à qui on a greffé des cellules tumorales au mois 0 (M0) et qui sont suivies pendant 12 mois. Les souris 1, 2, 4, 5 sont mortes respectivement à 3, 9, 5 et 2 mois. La souris 3 est toujours vivante à 12 mois, ses données sont donc *censurées* : on ne connaît pas sa durée de vie, on sait cependant qu'elle est supérieure à 12 mois.

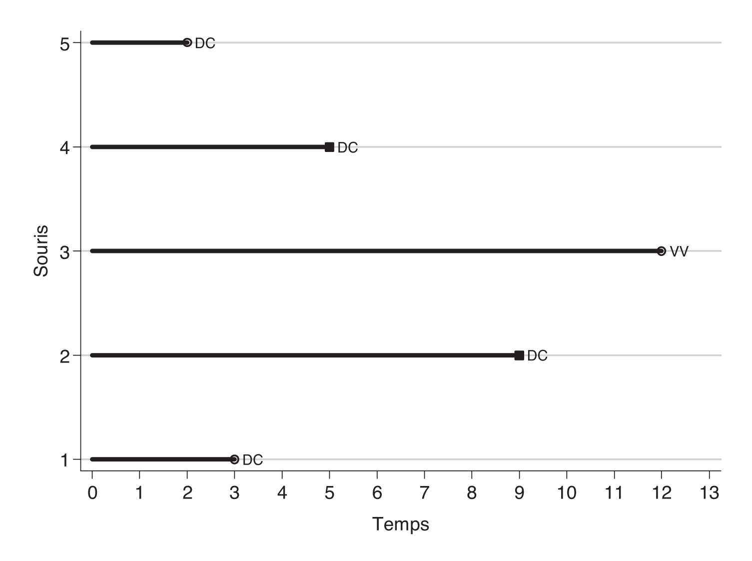


Figure 1. Représentation graphique de la durée de survie de 5 souris.
DC : décès ; VV : vivant.

Illustration de la notion de temps de participation

Dans un essai thérapeutique en médecine, l'inclusion des patients peut rarement se faire en même temps et les patients sont inclus dans l'étude au moment de leur diagnostic ou de l'initiation d'un nouveau traitement au cours de la période d'ouverture de l'essai (ici 12 mois). La date d'inclusion est donc échelonnée dans le temps. Dans la *figure 2*, les sujets 1 à 5 sont inclus dans l'étude en mai, avril, juin, mai et juillet respectivement. La date de point (*cf.* paragraphe « Définitions et notations ») est en décembre. Les sujets 5, 4 et 2 sont décédés en janvier de l'année suivante, août et septembre. Les données des deux sujets 1 et 3 sont censurées. En effet, le sujet 3 est toujours en vie à la date de fin d'étude et le sujet 1 a été perdu de vue en septembre. On note que la *figure 2* peut être translatée (même date d'origine) : c'est le délai entre la date d'inclusion d'un sujet et le décès qui est mesuré.

Définitions et notations

L'analyse de données de survie ne peut se faire sans la connaissance de l'événement d'intérêt, ni la connaissance de trois dates essentielles : la date d'origine, la date de l'événement (par ex., le décès) et la date de point.

• L'**événement** (EV) est une variable *binaire* dont les deux modalités sont oui (1 = événement) ou non (0 = pas d'événement). Pour le critère de survie globale, l'événement est le décès.

• **La date d'origine** (DO), définie dans le protocole, a la même définition (mais pas forcément une valeur) pour tous les sujets. C'est la date *à partir de laquelle chaque sujet commence à être surveillé pour détecter l'événement d'intérêt* ; elle correspond à la date de tirage au sort dans un essai thérapeutique, à la date de diagnostic (compte rendu d'anatomopathologie) dans une étude pronostique ou à la date de début du traitement dans une étude observationnelle.

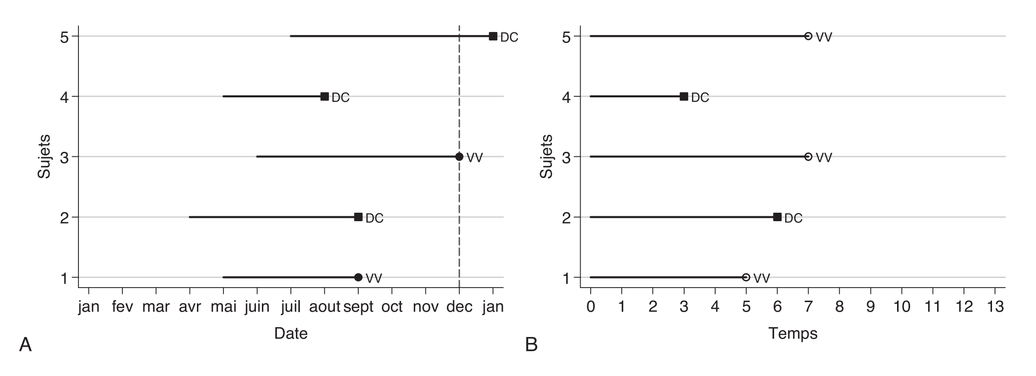


Figure 2. Représentation graphique de la durée de survie de 5 sujets en fonction du temps chronologique (A) et du temps de participation (B).
DC : décès ; VV : vivant.

• **La date de l'événement** (DEV) précise la date de l'événement pour les sujets présentant l'événement (*EV* = 1).

• **La date des dernières nouvelles** (DDN) est définie par la date de la dernière visite pour les sujets sans événement (EV = 0). Selon l'événement d'intérêt, les sujets sans événement sont considérés comme des censures à la *DDN*.

• **La date de point** (DP) est la date au-delà de laquelle on décide de ne plus comptabiliser les événements dans l'analyse, même si des événements sont observés au-delà de cette date. Toute l'information après cette date est ignorée. Son utilisation est conseillée dans la situation où il peut exister un suivi différent dans un des deux groupes à comparer, où au moment d'une analyse intermédiaire, par exemple.

• Le **recul** est défini par le délai entre la date de point et la date d'origine. Le recul permet de situer un sujet dans le temps par rapport à la date de point mais ne tient pas compte de l'occurrence ou non de l'événement d'intérêt.

• Le **suivi** est défini par le délai entre la date de point et la date des dernières nouvelles parmi les patients en vie. Le suivi permet d'évaluer la qualité du rythme de surveillance et la fréquence avec laquelle les sujets sont suivis.

Une fois les informations de délais et d'événements rassemblées, le **temps de participation (TP)** pour chaque sujet est défini par le délai entre la date d'origine et la date de l'événement (si EV = 1) ou par le délai entre la date d'origine et la date de point (si EV = 0). Son calcul dépend des situations suivantes :
– la date de dernières nouvelles est inférieure à la date de point :
  *$TP = DEV - DO$, si EV = 1 (l'événement a été observé),
  *$TP = DDN - DO$, si EV = 0 (l'événement n'a pas été observé) ;
– la date de dernières nouvelles est postérieure à la date de point :
  *$TP = DP - DO$ et EV = 0. Dans ce cas, pour l'analyse EV = 0.

Si l'on reprend les données présentées dans la *figure 2*, on observe que :
– le sujet 5 est décédé après la date de point, ainsi le décès n'est pas pris en compte comme événement et son temps de participation correspond alors à la différence entre les dates de point et d'origine ;
– le sujet 3 est vivant à la date de dernières nouvelles qui correspond à la date de point. Comme le premier sujet, son temps de participation correspond à la différence entre les dates de point et d'origine ;
– les sujets 2 et 4 sont décédés avant la date de point, ainsi leurs temps de participation correspondent à la différence entre leurs dates de décès et d'origine respectives ;
– Le sujet 1 a été perdu de vue (PDV) à la date de point – il n'a pas été revu en surveillance –, ainsi son temps de participation correspond à la différence entre les dates de dernières nouvelles et d'origine.

Hypothèses et conditions d'application des méthodes d'estimation et de comparaison

Dans ce chapitre, nous présentons différentes méthodes d'estimation et de comparaison de durées de survie. Toutes ces méthodes supposent que le processus de censure est non informatif, c'est-à-dire que la censure et le décès (ou un autre événement considéré) sont des événements indépendants, ou de manière équivalente : la censure à un instant t n'apporte pas d'information sur l'incidence d'événements ultérieurs. Cette indépendance entre le processus de censure et la survenue d'événements ultérieurs doit donc pouvoir être justifiée.

Estimation par la méthode de Kaplan-Meier

La variable « durée de vie » T, délai entre la date d'origine et la date du décès, est une variable aléatoire non négative considérée comme continue. Comme on s'intéresse de façon pragmatique à la probabilité de survivre au-delà d'un temps t, on définit la fonction de survie $S(t) = P(T \geq t)$ avec $S(0) = 1$ et $S(\infty) = 0$. Il existe plusieurs manières d'estimer et de comparer les fonctions de survie, mais les sections suivantes ne concernent que l'approche non paramétrique de Kaplan-Meier [2], la plus souvent utilisée en cancérologie.

Estimation de la probabilité de survie

La méthode de Kaplan-Meier repose sur une idée intuitive simple : « Être en vie après l'instant t, c'est avoir survécu jusqu'à l'instant t-1 et ne pas mourir à l'instant t (sachant que l'on était toujours vivant à l'instant t-1) ».

En langage statistique, « être en vie après l'instant t » correspond à $S(t) = P(T \geq t)$. De même, « être en vie jusqu'à l'instant $t-1$ » correspond à la probabilité de survie $S(t-1) = P(T \geq t-1)$. Enfin, « avoir survécu jusqu'à l'instant $t-1$ et ne pas mourir à l'instant t » correspond à la probabilité conditionnelle suivante : $q_t = P(T \geq t | T \geq t-1)$, c'est-à-dire être en vie après l'instant t sachant qu'on a survécu jusqu'à l'instant précédent. Par un processus itératif, on peut donc traduire l'affirmation initiale par l'égalité suivante :

$$\begin{aligned} S(t) &= P(T \geq t) \\ &= P(T \geq t | T \geq t-1) \times \Pr(T \geq t-1) \\ &= P(T \geq t | T \geq t-1) \times \ldots \times P(T \geq 1 | T \geq 0) \times P(T \geq 0) \end{aligned}$$

$$S(t) = q_t \times q_{t-1} \times \ldots \times 1 .$$

Soit d_i le nombre de décès observés au temps *i*, et n_i le nombre de sujets vivants juste avant temps *i*. Les probabilités instantanées de survie q_i et de décès $(1 - q_i)$ au temps *i* peuvent être estimées respectivement par $(n_i - d_i)/ n_i$ et d_i/n_i. Ainsi, l'estimateur de la survie au temps *t*, est donné par :

$$\hat{S}(t) = \prod_{i \leq t} \hat{q}_i$$

Si aucun décès n'est observé au temps *i*, alors $q_i = 1$ et l'estimateur de la survie *S(t)* est donc constant entre deux temps *i* consécutifs.

Exemple 1

Supposons que l'on souhaite estimer la survie d'un échantillon de 5 souris suite à l'injection d'une drogue. L'une des souris s'échappe de sa cage au jour 4 : ses données sont donc censurées au jour 4. Les 4 autres souris survivent 3, 6, 6 et 7 jours. En termes de taux bruts de survie, 20 % sont vivantes à 7 jours. La suite ordonnée des temps de participation des 5 souris est : 3, 4*, 6, 6, 7, où * correspond aux données censurées. On peut alors estimer les différentes probabilités en représentant les données dans un tableau *(tableau I)*.

Ainsi :

Tableau I. Construction de l'estimateur de Kaplan-Meier pour le groupe de 5 souris.

Temps (t)	d_i	n_i	d_i/n_i	$1 - (d_i)/n_i)$	*S(t)*
0	0	5	0	1	1
1	0	5	0	1	1
2	0	5	0	1	1
3	1	5	1/5	4/5	4/5
4	0	4	0	1	(4/5) x 1
5	0	4	0	1	(4/5) x 1
6	2	3	2/3	1/3	(4/5) x (1/3) = 4/15
7	1	1	1	0	(4/15) x 0 = 0

$$\hat{q}_1 = \hat{q}_2 = 1$$
$$\hat{q}_3 = (5-1)/5 = 4/5$$
$$\hat{q}_4 = (4-0)/4 = 1$$
$$\hat{q}_5 = (3-0)/3 = 1$$
$$\hat{q}_6 = (3-2)/3 = 1/3$$

La probabilité de survivre au-delà de 6 jours est donc estimée par :

$$\hat{S}(6) = \prod_{i \leq 6} \hat{q}_i = \hat{q}_1 \times \hat{q}_2 \times \hat{q}_3 \times \hat{q}_4 \times \hat{q}_5 \times \hat{q}_6 = 4/15$$

qui correspond à 27 %. On constate que seuls les instants auxquels au moins un décès est observé comptent effectivement dans le calcul.

Estimation de la variance de l'estimateur de Kaplan-Meier

Plusieurs estimateurs de la variance de l'estimateur de Kaplan-Meier ont été proposés, en particulier l'estimateur de Greenwood qui reste le plus fréquemment utilisé [3] et qui définit la variance de l'estimateur de Kaplan-Meier à un instant *t* comme suit :

$$\text{var}(\hat{S}(t)) = (\hat{S}(t)) \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}, \text{ où } \quad t_i \leq t \leq t_{i+1}.$$

Peto [4] a proposé d'estimer la variance à l'aide de la fonction suivante :

$$\text{var}(\hat{S}(t)) = (\hat{S}(t))^2 \frac{1 - \hat{S}(t)}{n_i}, \text{ où } \quad t_i \leq t \leq t_{i+1}.$$

La méthode de Peto est plus conservatrice, mais est plus appropriée pour les instants t où plus d'incertitude/variabilité subsiste, en particulier en fin de suivi où moins de sujets sont suivis.

Une fois la variance estimée, un intervalle de confiance (IC) peut être construit. En supposant que la distribution asymptotique de la fonction de survie suit une loi normale, l'IC à 95 % se calcule donc de la manière suivante : $S(t) \pm 1{,}96\ (Var[S(t)])^{1/2}$. Cependant, en présence de valeurs extrêmes, c'est-à-dire lorsque *S(t)* s'approche de 0 ou 1, cette formule peut conduire à un IC contenant des valeurs négatives ou au-delà de 1. Rothman a donc proposé [5] la formule suivante :

$$IC\ 95\% = \frac{M}{M + z^2_{1-\alpha/2}} \left[\hat{S}(t) + \frac{z^2_{1-\alpha/2}}{2M} \pm z_{1-\alpha/2} \sqrt{Var\left[\hat{S}(t)\right] + \frac{z^2_{1-\alpha/2}}{4M^2}} \right],$$

avec

$$M = \frac{\hat{S}(t)\left[1 - \hat{S}(t)\right]}{Var\left[\hat{S}(t)\right]}$$

et S(t) correspond à la variance telle que définie par Greenwood.

Exemple 2

Considérons une cohorte de 20 sujets inclus dans une étude thérapeutique. Deux traitements, A et B, sont évalués. Le critère de jugement est la survie globale, c'est-à-dire le délai de survenue du décès. Le *tableau II* représente les temps de participation (TP) en jours ordonnés des 20 sujets pour chacun des deux groupes de traitements. On utilise l'astérisque * pour des données censurées. À défaut, on considère donc que l'événement d'intérêt, ici le décès, a été observé.

On pourrait ne s'intéresser qu'aux sujets pour lesquels un décès a été effectivement observé et ne reporter qu'une proportion brute. Cette approche entraînerait cependant une perte d'information et biaiserait les estimations. En effet, il faut intégrer les données censurées dans l'analyse car pour ces sujets, même si on ne connaît pas leur délai de survie exacte, on connaît la durée pendant laquelle ils ont survécu.

Estimation de la fonction de survie

Les calculs nécessaires à l'estimation de la fonction de survie ne sont effectués qu'aux instants de survenue d'un décès et sont présentés dans le *tableau III* pour le groupe A.

L'un des sujets est perdu de vue au 88e jour et ainsi participe au calcul uniquement jusqu'à l'instant 88 (délai de censure). Il compte donc dans les sujets exposés au risque de décès jusqu'à cet instant-là. Les calculs peuvent être réalisés de la même manière pour le groupe B.

Il est à noter que si un décès et une censure se produisent au même instant, on considère dans les calculs que la censure se produit après le décès.

Dans le groupe A, il y a 8 décès sur 10 et on aurait pu résumer les données par une probabilité brute de survie estimée à 20 %. Cela aurait été le cas si les sujets vivants à 88 et 111 jours avaient survécu au-delà du dernier décès observé. Mais les probabilités de survie varient dans le temps et la probabilité de survie à 105 jours est égale à 37,5 %, et à 192 jours, elle est égale à 0 %, car le dernier sujet est décédé.

Tableau II. Temps de participation par groupe.

Groupe	Temps de participation (jours)
A	24, 49, 50, 65, 88, 88*, 105, 111*, 125, 192
B	102, 103, 118*, 134*, 160, 202*, 204, 209, 235, 257

Tableau III. Construction de l'estimateur de Kaplan-Meier et de sa variance estimée pour le groupe A des sujets de l'exemple 2.							
TP mois	**EV**	**Intervalle [ti, ti+1[**	n_i	d_i	*qi*	*S(t)*	*Var(S(t))*
0		[0, 24[	10	0	1	**1**	–
24	DCD	[24, 49[	10	1	9/10	1 x 9/10 = 9/10 = **0,900**	$(9/10)^2$ x 1/90 = **0,009**
49	DCD	[49, 50[	9	1	8/9	9/10 x 8/9 = 8/10 **= 0,800**	$(8/10)^2$ x (1/90+1/72) = **0,016**
50	DCD	[50, 65[	8	1	7/8	8/10 x 7/8 =7/10 **= 0,700**	$(7/10)^2$ x (1/90+1/72+1/56) = **0,021**
65	DCD	[65, 88[	7	1	6/7	7/10 x 6/7 = 6/10 **= 0,600**	$(6/10)^2$ x (1/90+1/72+1/56+1/42) = **0,024**
88	DCD	[88, 88[	6	1	5/6	6/10 x 5/6 = 5/10 = **0,500**	$(5/10)^2$ x (1/90+1/72+1/56+1/42 +1/30) = **0,025**
88	VV	[88, 105[	5	0	5/5	5/10 x 5/5 = 5/10 = **0,500**	-
105	DCD	[105, 111[	4	1	3/4	5/10 x 3/4 = 3/8 = **0,375**	$(3/8)^2$ x (1/90+1/72+1/56+1/42 +1/30+1/12) = **0,026**
111	VV	[111, 125[	3	0	3/3	3/8 x 3/3 = 3/8 = **0,375**	-
125	DCD	[125, 192[	2	1	1/2	3/8 x 1/2 = 3/16 = 0,1875	$(3/16)^2$ x (1/90+1/72+1/56+1/42 +1/30+1/12+1/2) = **0,024**
192	DCD	[192, ...[	1	1	1/1	0	–

TP : temps de participation ; EV : événement ; DCD : décédé ; VV : vivant.

Estimation des intervalles de confiance

Par la formule symétrique, on obtient l'*IC95 %* à 3 mois : [0,19 - 0,81]. En appliquant la formule de Rothman, on obtient M = 10, et l'*IC95 %* à 3 mois : [0,237 - 0,763].

Courbe de survie

La courbe de survie doit comporter un titre, des légendes et un IC à 95 % (ou à défaut le nombre de sujets à risque). La *figure 3* est une représentation graphique de la survie globale pour les sujets du groupe A, les barres verticales représentant les IC à 95 % avec le nombre de sujets à risque aux intervalles réguliers en bas de la courbe.

Autres méthodes

Méthode actuarielle

Cette méthode, aujourd'hui peu utilisée depuis l'arrivée de l'informatique car moins précise, était initialement appréciée pour sa moindre lourdeur de calcul dans les études avec un grand nombre d'événements. Les intervalles de temps ne sont pas déterminés par la survenue de chaque décès mais sont *fixés a priori* et on situe la fin de l'histoire du sujet dans l'un de ces intervalles selon son temps de participation. Son utilisation pourrait être limitée aux études avec des suivis réguliers organisés (diagnostic d'intervalle).

Calcul direct

Intuitivement, on pourrait utiliser une méthode « flash » qui permet d'estimer la survie à un instant donné, choisi *a priori*. Mais ne sont pris en compte dans le calcul que les sujets ayant un recul suffisant. Cette dernière méthode a comme inconvénient majeur d'éliminer des calculs tous

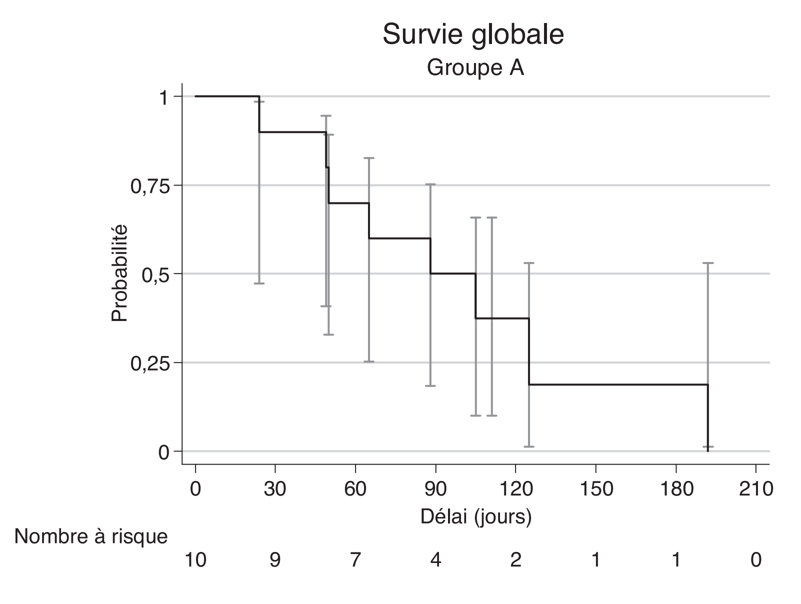


Figure 3. Survie globale du groupe A (intervalles de confiance à 95 %).

les patients n'ayant pas un recul suffisant et ceux perdus de vue. Cette méthode ne doit pas en outre être utilisée pour donner des estimations successives et en déduire une représentation graphique. Le mode de calcul peut aussi amener à d'apparentes contradictions. On préférera donc utiliser les méthodes de Kaplan-Meier (ou actuarielle) qui prennent en compte la totalité du temps de participation de chaque sujet.

Comparaison de courbes par le test du log-rank

D'un point de vue statistique, le problème posé est la comparaison de deux échantillons de patients ou plus en termes de durées de survie éventuellement censurées. L'hypothèse nulle à tester est celle de l'identité des distributions de survie dans les échantillons à comparer. Quatre tests peuvent être utilisés pour comparer les taux de survie entre les groupes : le test du log-rank, le test de Gehan, le test de Tarone-Ware et le test de Peto-Prentice. Ces méthodes reposent sur une approche conditionnelle, à savoir le temps où des décès surviennent est supposé fixé et on compare le nombre de décès observés dans chaque groupe à son espérance calculée sous l'hypothèse nulle H_0 d'identité des distributions de survie.

Notations et statistique de test

On se limite dans un premier temps à la comparaison de deux groupes A et B. Les temps observés de décès ordonnés des deux échantillons sont notés t_1, t_2, ..., t_K. En t_i, les nombres de décès observés dans chacun des groupes A et B sont notés :

m_{Ai} et m_{Bi} avec $m_{Ai} + m_{Bi} = m_i$ et $m_i > 0$

Les nombres de sujets exposés au risque de décès en cet instant sont :

n_{Ai} et n_{Bi} avec $n_{Ai} + n_{Bi} = n_i$

Les nombres de décès observés dans chaque groupe sont notés respectivement O_A et O_B.

$$O_A = \sum_{i=1}^{k} m_{A_i} \qquad O_B = \sum_{i=1}^{k} m_{B_i}$$

Les observations peuvent être résumées sous la forme d'un tableau 2*2 *(tableau IV)* pour chacun des temps de décès t_i.

L'hypothèse nulle H_0 à tester est celle de l'égalité des fonctions de survie dans les groupes A et B. Donc, sous H_0, au temps de décès t_i, la proportion attendue de décès parmi les sujets à risque est identique dans les deux groupes. La statistique de test s'écrit :

$$L_W = \frac{\left[\sum_{i=1}^{k} w_i \left(m_{B_i} - m_i \frac{n_{B_i}}{n_i} \right)\right]^2}{\sum_{i=1}^{k} w_i^2 m_i \frac{(n_i - m_i)}{(n_i - 1)} \frac{n_{A_i} n_{B_i}}{n_i^2}}$$

avec w_i représentant le poids attribué à chaque observation (*cf*. paragraphe suivant). Cette statistique suit asymptotiquement sous H_0 une loi de χ^2 à un degré de liberté.

Tableau IV. Nombre de sujets décédés et toujours en vie à l'instant t_i par groupe.

	Décédés	**Vivants après t_i**	**Total**
Groupe A	m_{Ai}	n_{Ai} - m_{Ai}	n_{Ai}
Groupe B	m_{Bi}	n_{bi} - m_{bi}	n_{Bi}
Total	m_i	n_i - m_i	n_i

Test du log-rank

Le test du log-rank correspond à la pondération la plus simple ($w_i = 1$ pour tout i), qui attribue à chaque décès le même poids, quel que soit l'instant de survenue. Le test revient à comparer le nombre O_B de décès observés dans le groupe B au nombre E_B de décès attendus sous H_0 dans ce groupe, avec

$$E_B = \sum_{i=1}^{k} e_{B_i} \quad \text{et} \quad e_{B_i} = m_i n_{B_i} / n_i$$

Ce test est connu sous le nom de Mantel-Haenzel ou du log-rank. La statistique de test est obtenue en appliquant la pondération $w_i = 1$ à la formule ci-dessus. Le test du log-rank est le test standard de comparaison de deux courbes de survie. Lorsque le résultat est significatif, il permet de rejeter l'hypothèse que les deux courbes proviennent de la même population. Il analyse si globalement, au niveau de chaque décès, la distance entre les deux courbes est plus grande que ce que pourrait expliquer le hasard. Ainsi, le test du log-rank analyse les courbes dans leur globalité.

Cependant, pour réaliser un test du log-rank, il est nécessaire de vérifier une hypothèse appelée l'hypothèse des risques proportionnels : au cours du temps, l'écart entre les courbes de survie (plus spécifiquement sur l'échelle logarithmique) de deux groupes reste constant au cours du temps. Une façon graphique de vérifier la validité de cette hypothèse est de représenter l'estimation de la fonction de risque cumulée H(t) [ou son équivalent –log(S(t)] en fonction du temps pour chacun des groupes. L'hypothèse des risques proportionnels est graphiquement acceptable

si les courbes obtenues sont parallèles. Des méthodes statistiques permettant de tester l'hypothèse de risques proportionnels existent et sont programmées dans de nombreux logiciels. L'analyse visuelle des courbes doit donc toujours accompagner l'interprétation d'un test du log-rank.

Autres tests : Gehan, Tarone-Ware et Peto-Prentice

Le test de Gehan (appelé aussi test de Wilcoxon ou de Breslow), le test de Tarone-Ware et le test de Peto-Prentice reposent sur le même principe que le test du log-rank, à la différence que les poids w_i en t_i sont différents de 1 :

$w_i = 1$ pour le test L_L log-rank [4, 6]

$w_i = n_i$ pour le test L_G [7]

$w_i = ni^{1/2}$ pour le test L_{TW} [8]

$w_i = S_i*$ qui pour le test de L_P [9]

avec

$$S_i^* = \prod_{j=1}^{i} \frac{n_j}{n_j + m_j}.$$

Cette quantité est proche de l'estimateur de Kaplan-Meier de la fonction de survie qui, sous H_0, est supposée commune aux deux échantillons.

Le test du log-rank est souvent plus approprié quand l'alternative de l'hypothèse nulle d'égalité des survies entre les groupes respecte le principe des risques proportionnels entre les groupes à chaque temps (chapitre IV.2 « Modèle de Cox et index pronostique », page 213). Cette hypothèse devrait donc être vérifiée lors de la réalisation de chaque test.

Les autres tests sont plus aptes à déceler une différence entre les groupes en présence de nombreux décès précoces. Par ailleurs, la comparaison des pondérations utilisées montre que la statistique de Gehan dépend davantage de la distribution des censures que la statistique de Peto-Prentice.

En conclusion, le test du log-rank est le test le plus employé. Cependant, l'interprétation des résultats des tests doit prendre en considération la taille de l'effectif étudié ainsi que le profil et la distribution des censures. Quand le nombre de sujets est faible, les résultats des tests doivent être interprétés avec prudence. Le *tableau V* présente pour chaque test étudié les caractéristiques, avantages et inconvénients.

Remarque : Il est également possible d'étendre chacun des tests dans le cas de la comparaison de plus de deux groupes et dans le cas d'une stratification sur un facteur pronostique.

Tableau V. Caractéristiques, avantages et limites des tests de comparaison de courbes de survie.			
Nom du test	**Caractéristiques**	**Avantages**	**Inconvénients**
Log-rank	Même pondération pour chaque sujet	Considéré comme le test de référence lorsque l'hypothèse des risques proportionnels est vérifiée	
Gehan (ou Wilcoxon ou Breslow)	Pondération = nombre de sujets exposés au risque à chaque temps. Les décès précoces ont un poids plus élevé que les décès tardifs	Apte à déceler une différence en cas de décès nombreux en phase précoce de l'étude	Statistique fortement dépendante de la distribution des sujets censurés Moins puissant que le test du log-rank mais peut être utilisé pour montrer une différence entre les courbes de survie portant essentiellement sur les survies de courte durée
Tarone-Ware	Pondération = racine carrée du nombre de sujets exposés au risque à chaque temps	Idem que Gehan	Idem que Gehan, mais moins dépendant de la distribution des censures
Peto-Prentice	Pondération = proche de l'estimateur de Kaplan-Meier Les décès précoces ont un poids plus élevé que les tardifs	Idem que Gehan	Statistique dépendante de la distribution des sujets censurés

Exemple 2 (suite)

Reprenons l'exemple précédent. La courbe de Kaplan-Meier du groupe A a été obtenue précédemment *(figure 3)*. Les deux courbes peuvent être représentées simultanément *(figure 4)*. On s'intéresse à la comparaison des deux courbes. De la même façon que pour le groupe A, on estime tout d'abord la probabilité de survie pour le groupe B *(tableau VI)*.

On calcule ensuite la statistique du log-rank tel qu'illustré dans le *tableau VII*. Cet exercice est fastidieux sur des données réelles, mais il existe des logiciels spécifiques (*cf*. chapitre VI.5 « Les logiciels », page 370) permettant ce calcul.

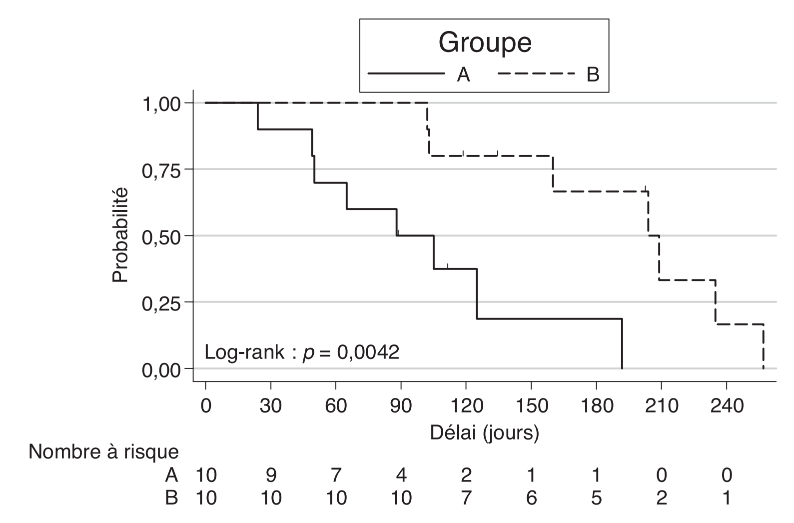


Figure 4. Courbes de survie des groupes A et B pour les 20 sujets de l'exemple 2.

Tableau VI. Calcul détaillé de la survie globale dans le groupe B.							
TP en mois	**État**	**Intervalle [ti, ti+1[**	n_i	d_i	$\hat{q}i$	*S(t)*	*Var(S(t))*
0		[0, 102[	10	0	1	**1**	–
102	DCD	[102, 103[	10	1	9/10	**0,900**	0,009
103	DCD	[103, 118[	9	1	8/9	**0,800**	0,016
118	VV	[118, 134[	8	0	8/8	**0,800**	–
134	VV	[134, 160[	7	0	7/7	**0,800**	–
160	DCD	[160, 202[	6	1	5/6	**0,667**	0,026
202	VV	[202, 204[	5	0	5/5	**0,667**	–
204	DCD	[204, 209[	4	1	3/4	**0,500**	0,035
209	DCD	[209, 235[	3	1	2/3	**0,333**	0,034
235	DCD	[235, 257[	2	1	1/2	**0,167**	0,022
257	DCD	[257, ...[	1	1	1/1	**0**	–

TP : temps de participation ; DCD : décédé ; VV : vivant.

Tableau VII. Calcul de la statistique du log-rank afin de comparer la survie du groupe A à celle du groupe B pour les 20 sujets de l'exemple 2.

TP jours	Groupe	État	Décès observés			Nb de patients exposés au risque			Décès attendus	
			O_{Ai}	O_{Bi}	m_i	n_{Ai}	n_{Bi}	n_i	$E_{Ai} = m_i \times (n_{Ai}/n_i)$	$E_{Bi} = m_i \times (n_{Bi}/n_i)$
24	A	DCD	1	0	1	10	10	20	1 x (10/20) = 0,50	1 x (10/20) = 0,50
49	A	DCD	1	0	1	9	10	19	1 x (9/19) = 0,474	1 x (10/19) = 0,526
50	A	DCD	1	0	1	8	10	18	1 x (8/18) = 0,444	1 x (10/18) = 0,556
65	A	DCD	1	0	1	7	10	17	1 x (7/17) = 0,412	1 x (10/17) = 0,588
88	A	DCD	1	0	1	6	10	16	1 x (6/16) = 0,375	1 x (10/16) = 0,625
88	A	VV	0	0	0	5	10	15	0	0
102	B	DCD	0	1	1	4	10	14	1 x (4/14) = 0,286	1 x (10/14) = 0,714
103	B	DCD	0	1	1	4	9	13	1 x (4/13) = 0,308	1 x (9/13) = 0,692
105	A	DCD	1	0	1	4	8	12	1 x (4/12) = 0,333	1 x (8/12) = 0,667
111	A	VV	0	0	0	3	8	11	0	0
118	B	VV	0	0	0	2	8	10	0	0
125	A	DCD	1	0	1	2	7	9	1 x (2/9) = 0,222	1 x (7/9) = 0,778
134	B	VV	0	0	0	1	7	8	0	0
160	B	DCD	0	1	1	1	6	7	1 x (1/7) = 0,143	1 x (6/7) = 0,857
192	A	DCD	1	0	1	1	5	6	1 x (1/6) = 0,167	1 x (5/6) = 0,833
202	B	VV	0	0	0	0	5	5	0	0
204	B	DCD	0	1	1	0	4	4	0	1
209	B	DCD	0	1	1	0	3	3	0	1
235	B	DCD	0	1	1	0	2	2	0	1
257	B	DCD	0	1	1	0	1	1	0	1
Total			8	7	15				E_A = 3,664	E_B = 11,336

TP : temps de participation ; DCD : décédé ; VV : vivant.

On vérifie que la somme des décès observés dans les deux groupes ($O_A + O_B$) correspond bien à la somme des décès attendus ($E_A + E_B$). Les quatre statistiques permettant la comparaison des groupes peuvent alors être calculées selon les formules énumérées précédemment et comparées à une distribution de chi-2 à un degré de liberté :

- log-rank : (8,19, $p = 0{,}0042$) ;
- Breslow : (6,98, $p = 0{,}0082$) ;
- Tarone-Ware : (7,56, $p = 0{,}0060$) ;
- Peto : (7,37, $p = 0{,}0066$).

Représentation graphique

La probabilité de survie peut être représentée en fonction du temps, on parle alors de courbe de survie. Inversement, la courbe correspondant à la proportion cumulée de patients présentant l'événement, c'est-à-dire la courbe d'incidence cumulée, peut également être représentée, dépendant de l'intérêt principal. Pocock [10] a défini les bonnes pratiques de représentation graphique des données de survie.

Représentation de la survie ou de l'incidence cumulée ?

Si le taux d'événement est faible (moins de 30 %), la représentation de l'incidence cumulée sera plus claire. L'introduction de l'axe entier 0-100 % pourra nuire à la lisibilité, en particulier si la proportion d'événements est faible. De façon systématique, il est préférable d'éviter une rupture dans l'axe des ordonnées.

Quelle étendue temporelle de la courbe de survie ?

Faut-il que l'axe des abscisses couvre tous les temps de participation ou doit-on représenter la courbe de survie pour un délai écourté ? Du fait de la tendance naturelle de l'œil à se focaliser sur la partie droite de la courbe qui est la région de plus forte incertitude du fait du nombre de patients restant à risque souvent limité, Pocock conseille de ne poursuivre la courbe que tant que ce nombre reste supérieur à 10-20 % de l'effectif initial afin de ne pas biaiser l'interprétation des résultats (par ex., en cas d'un échantillon de 100 patients, la courbe s'arrêterait dès lors qu'il resterait moins de 10 patients à risque).

De plus, il est habituel qu'une courbe s'aplanisse après un certain délai lorsque la survenue des événements est moins fréquente. Il n'est alors pas licite d'interpréter cet aplanissement comme une stabilité de la fonction de survie mais comme un manque d'information liée à des sujets tous censurés sauf si le *nombre de sujets encore à risque* reste encore important. À l'inverse, si la dernière donnée est un décès, la courbe de survie plonge vers l'axe des abscisses. Cela ne signifie pas non plus qu'aucun sujet ne survivrait au-delà de ce temps de suivi. C'est pourquoi on

recommande fortement dans les publications de faire figurer, en dessous des courbes de survie, le nombre de sujets toujours à risque à différents temps pour évaluer rapidement la précision de la courbe à ces temps.

Le degré d'incertitude statistique des estimateurs

Une notion de la précision des estimations doit figurer car l'impression visuelle peut parfois apparaître bien plus convaincante que ne l'est la réalité. Il est donc possible, soit d'afficher à certains délais d'intérêt une représentation de l'erreur standard (ou de l'IC à 95 %), soit d'indiquer une estimation globale de la différence entre les groupes par le risque relatif, par exemple, en association avec son IC 95 % et/ou la valeur du *p* du test du log-rank.

Résumer les données

La courbe de survie permet de visualiser la survie en fonction du temps. Dans certains cas, notamment en vue de comparaisons entre groupes distincts, on souhaite résumer les informations de survie sous la forme d'une donnée numérique. Dans ce contexte, une erreur habituelle consiste à résumer les données de survie par la proportion de sujets encore vivants à un moment donné après le début d'une étude (*cf.* calcul direct). À moins que les temps de participation de chaque sujet soient identiques, cette approche sera biaisée par manque de recul suffisant.

Plutôt que de reporter le nombre de sujets toujours en vie en fin de suivi, on peut s'intéresser à la durée de vie ou au temps de participation. Reporter la durée de vie moyenne n'est approprié qu'en l'absence de données censurées. En la présence de données censurées, c'est en fait la médiane qu'il faut reporter (et non la moyenne) : il s'agit de la durée nécessaire afin que le taux de survie estimé par la méthode de Kaplan-Meier atteigne 50 %. Mais ce n'est qu'un point sur la courbe tout comme une probabilité de survie à un temps donné.

La médiane de survie permet de résumer la rapidité de décroissance de la fonction de survie et donc de quantifier la gravité de la maladie si l'événement mesuré est le décès. La *figure* 5 représente deux courbes de survie hypothétiques (médianes de 8 et 15 mois) où l'imprécision de l'estimation de la médiane de survie est présentée lorsque peu d'événements surviennent aux alentours de celle-ci.

Une limite du calcul de la médiane de survie est donc l'absence de prise en compte de la pente de la courbe de survie après 50 %. C'est dans ce contexte que certains auteurs comme Messori [11] ont proposé une méthode pour améliorer la description intégrale de la courbe. Le calcul de l'aire sous la courbe permet de mesurer la moyenne du temps de survie avec une extrapolation de la fin de la courbe par le modèle de Gompertz. D'autres auteurs comme Vaidya [12] ont proposé la mesure non pas du temps de survie brute mais la proportion de vie restante par rapport à l'espérance de vie selon l'âge. Ces notions sont intéressantes mais elles nécessitent des hypothèses supplémentaires.

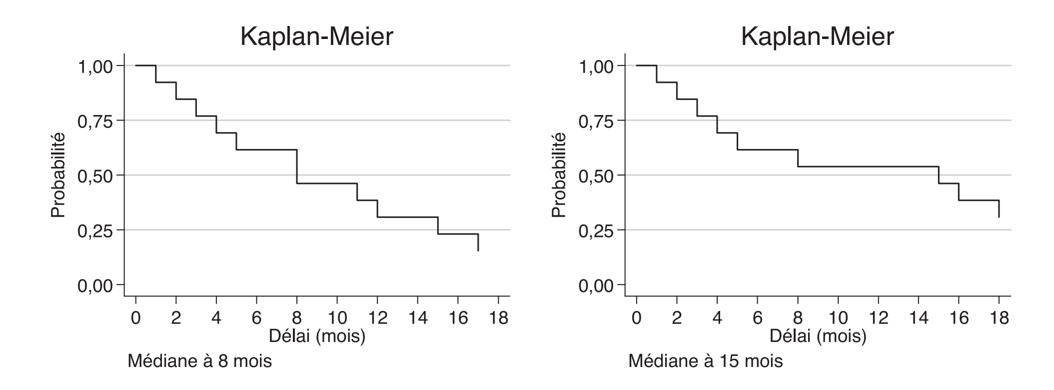


Figure 5. Exemple d'imprécision dans le calcul de la survie médiane.

Le suivi

Les estimations de probabilités de survie dépendent des temps de participation des sujets. Il est donc nécessaire de préciser la durée de suivi lorsque l'on reporte les résultats en termes de probabilité de survie. Le suivi médian est souvent utilisé pour indiquer la qualité du suivi des patients en vie, car la validité des estimations en dépend. Il est calculé par la méthode de Schemper [13]. Il est généralement admis que les estimations au-delà du suivi médian ne sont pas valides et donc un suivi aussi complet que possible est primordial notamment dans les essais cliniques. Les perdus de vue sont souvent définis comme les sujets ayant un recul suffisant, mais pour lesquels on est sans nouvelles depuis un certain nombre d'années. Il convient de limiter leur nombre (< 10-15 %) et de vérifier par une enquête supplémentaire sur un sous-échantillon que cette assimilation n'est pas trop mauvaise. Les sujets perdus de vue peuvent donc être la source d'un biais de sélection qui sera d'autant plus présent que les perdus de vue seront nombreux. Clark [14] a proposé un index, l'index C, permettant de quantifier le *follow-up* en résumant le temps perdu de suivi (*cf.* chapitre III.4 « Suivi et surveillance », page 181).

Points importants

Hormis les points déjà abordés, il faut insister sur la nécessité de vérifier les points suivants :
- l'absence de censure informative : les perdus de vue doivent être comparables aux non perdus de vue ;
- un suivi suffisamment long et suffisamment exhaustif en complément du simple décompte des perdus de vue ;
- l'homogénéité des caractéristiques des patients au cours de la période (comparabilité des patients de début et de fin de période) ;
- la représentativité de l'échantillon pour l'extrapolation des résultats ;
- un fichier de données qui doivent avoir été vérifiées avant toute analyse (*cf.* chapitre VI.1 « Gestion des données », page 331) ;
- une prise en charge de données, notamment celles concernant l'évolution (temps de participation et événement), qui doit être identique pour tous les patients :

– rechercher avec la même ardeur à récupérer les dates des dernières nouvelles que le patient soit connu comme décédé ou non,
– et ne pas prendre en compte les événements survenus après la date de point.

Références

1. Graunt J. *Natural and political observations made upon the bills of mortality*. London: Thos. Roycroft, 1662.

2. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc* 1958 ; 53 : 457-81.

3. Greenwood Major. A report on the natural duration of cancer. In : *Reports on Public Health and Medical Subjects, vol. 33* London: His Majesty's Stationery Office, 1926 ; 1-26.

4. Peto R, Pike MC, Armitage P, *et al*. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. II. Analysis and examples. *British Journal of Cancer* 1977 ; 35 (1) : 1-39.

5. Rothman KJ. Estimation of confidence limits for the cumulative probability of survival in life table analysis. *J Chron Dis* 1978 ; 31 : 557-60.

6. Bland JM, Altman DA. The logrank test. *BMJ* 2004 : 328 : 1073.

7. Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika* 1965 ; 52 : 203-23.

8. Tarone RE. Tests for trend in life table analysis. *Biometrika* 1975 ; 62 : 679-82.

9. Prentice RL. Linear rank tests with right censored data. *Biometrika* 1978 ; 65 : 167-79.

10. Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: Good practice and pitfalls. *Lancet* 2002 (359) : 1686-9.

11. Messori A, Becagli P, Trippoli S. Median *versus* mean lifetime survival in the analysis of survival data. *Haematologica* 1997 ; 82 (6) : 730.

12. Vaidya JS, Mitra I. Fraction of normal remaining life span: A new method for expressing survival in cancer. *Br Med J* 1997 ; 14 : 1682-4.

13. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Controlled Clinical Trials* 1996 ; 17 : 343-6.

14. Clark TG, Altman DG, De Stalova BL. Quantification of the completeness of follow-up. *Lancet* 2002 ; 359 : 1309-10.

Facteurs pronostiques et prédictifs de la réponse à un traitement

X. Paoletti, S. Mathoulin-Pélissier, S. Michiels

Qu'adviendra-t-il à un patient est une question de prédiction individuelle, mais pour le praticien ou pour le statisticien, la question est plutôt de connaître les caractéristiques qui sont associées au devenir du patient et comment les combiner pour être capable de constituer des groupes homogènes quant à leur risque de développer un événement d'intérêt. La recherche de facteurs pronostiques occupe une place importante dans la littérature médicale et épidémiologique. Elle permet d'identifier des populations homogènes pour les essais cliniques, d'adapter les rythmes de surveillance, d'identifier des populations à risque pour des traitements préventifs ou pour un dépistage. Ainsi, une amniocentèse pour le dépistage d'une trisomie 21 chez le fœtus ne sera-t-elle effectuée que lorsque le dosage de l'hormone HCG est jugé anormal.

En cancérologie, l'événement à prédire est souvent le décès ou la récidive tumorale, mais toute variable « à expliquer » peut être d'intérêt. En particulier, la réponse à un traitement, que ce soit la réponse clinique ou la toxicité associée, occupe une place particulière, à savoir identifier des groupes de patients pour lesquels les traitements sont actifs. Les traitements en cancérologie bénéficient seulement à une fraction de la population. Mais est-il possible d'identifier cette fraction de patients avant la mise sous traitement à partir de caractéristiques histopathologiques, biologiques ou génétiques d'une tumeur ? Il s'agit ici de la recherche de facteurs prédictifs. De nombreuses confusions seraient évitées si on précisait « prédictif de la réponse au traitement T ». Cette thématique largement soutenue par la recherche translationnelle et la compréhension de mécanismes d'action des nouvelles molécules a connu un essor sans précédent avec l'arrivée des traitements moléculaires ciblés ; en effet, leurs mécanismes d'action requièrent souvent la présence d'une ou de plusieurs cibles données si bien que le traitement de l'ensemble de la population conduira à sous-estimer, voire à ne pas détecter, un éventuel bénéfice, tout en exposant inutilement des patients à des toxicités souvent graves ; en outre, ces traitements ont des coûts financiers tels que la société veut restreindre leur utilisation aux seuls patients bénéficiaires ; certains parleront alors à ce propos de médecine stratifiée ou même personnalisée.

Dans ce chapitre, nous traiterons essentiellement des méthodes statistiques et des plans d'expérience pour rechercher et valider des facteurs pronostiques et prédictifs de la réponse à un traitement. Après avoir précisé la différence entre variable pronostique et variable prédictive, nous explorerons les facteurs pronostiques en présentant la construction des scores de risque et des scores de prédiction individuels, les mesures pour évaluer leur qualité en distinguant la

discrimination et la calibration et, enfin, les techniques de validation. Dans la dernière partie, nous présenterons les plans d'expérience et tests pour l'identification et la validation de scores prédictifs de la réponse à un traitement donné.

Variables pronostiques et variables prédictives de la réponse à un traitement donné

Une variable pronostique est typiquement mesurée à l'origine, que ce soit au diagnostic, à la randomisation ou à tout point marquant le début du suivi. Elle prédit un événement futur, indépendamment du fait que les patients soient ou pas traités.

Une variable prédictive de la réponse au traitement est un cas particulier de variable pronostique. C'est une variable d'interaction qui modifie l'effet d'un traitement ou d'une exposition sur un événement futur. En épidémiologie, on préférera le terme de « modificateur d'effet ».

À titre d'illustration, nous vous proposons d'observer des résultats à partir d'une étude de simulation sur un échantillon de 800 patients. Cet échantillon a été généré en supposant 2 ans de recrutement et 5 ans de suivi, une distribution de survie exponentielle, une prévalence de 50 % du facteur prédictif et des effets du traitement d'un *hazard ratio* (HR) égal à 0,5 et 1 dans chaque strate. Le taux de survie dans le bras contrôle est de 40 % et de 62 % dans la population marqueur positif et marqueur négatif.

Un facteur pronostique n'est pas forcément prédictif de la réponse au traitement mais une variable prédictive doit être pronostique dans au moins un des bras de traitement. La *figure 1* illustre cela : sur la partie inférieure, on voit que le pronostic des patients est moins bon chez les patients marqueurs positifs uniquement lorsqu'ils sont traités par le contrôle. Les impacts pronostiques et prédictifs, s'ils sont conjoints, peuvent être de sens opposé. Par exemple, la surexpression des récepteurs HER-2 pour les tumeurs du sein est associée à un mauvais pronostic mais c'est un facteur prédictif fort du bénéfice de l'Herceptin®.

Les méthodes statistiques en découlent naturellement. Dans le premier cas, il s'agira de montrer une association entre deux variables et, dans le second, une interaction entre l'effet d'un traitement et la variable prédictive sur le devenir.

Il faut noter qu'il existe un abus de langage puisque le plus souvent sont identifiés des facteurs prédictifs de la *non*-réponse au traitement. Que ce soit le statut HER-2 ou celui de la mutation du gène *KRAS* dans le cancer du côlon, les valeurs négatives du premier et la mutation de l'autre prédisent avec une grande certitude une absence de bénéfice du traitement. Par analogie avec les critères diagnostiques, on pourrait dire qu'ils ont une bonne valeur prédictive négative.

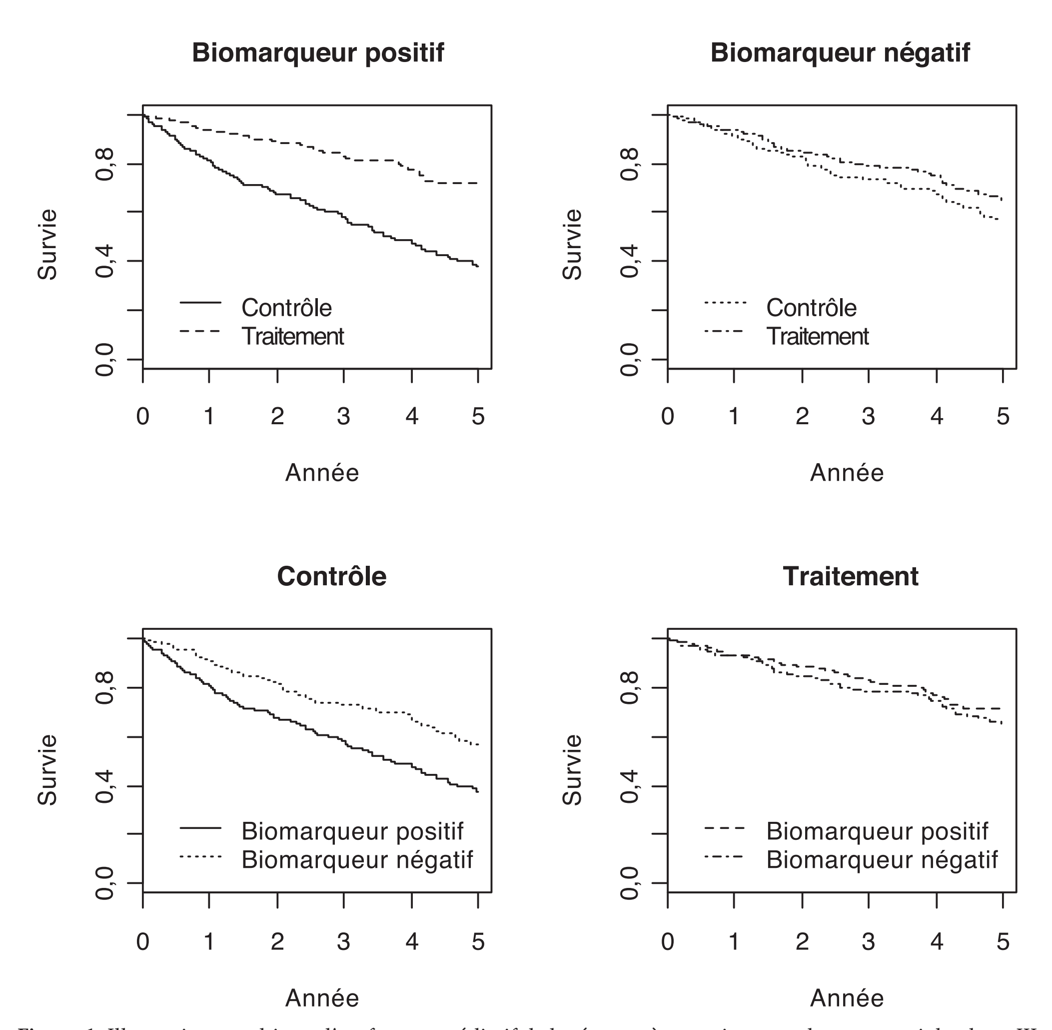


Figure 1. Illustration graphique d'un facteur prédictif de la réponse à un traitement dans un essai de phase III fictif.
Un facteur prédictif peut être représenté en comparant les courbes de survie du traitement avec le contrôle pour chaque valeur du facteur (figures supérieures) ou en comparant l'impact pronostique du marqueur dans les bras traitement et contrôle (figures inférieures).

Facteurs pronostiques

Construction de score de prédictions

Soit Y la variable à expliquer et $X_1, \ldots, X_K$, K facteurs pronostiques de Y. Y pourra être une variable binaire, continue, de survie, de comptage, etc.

La construction du score de prédiction individuelle revient à la construction d'un modèle paramétrique. Sans entrer dans le détail des difficultés de la construction des modèles, on en rappelle les grandes étapes :

- transformation des variables ;
- choix d'une forme fonctionnelle d'un modèle (additif, multiplicatif, linéaire, non linéaire, etc.) ;
- gestion des données manquantes ;
- sélection des variables à inclure dans le modèle, en prévoyant ou pas une procédure de présélection (analyse univariée, analyse en composantes principales). Des techniques alternatives comme les arbres de régression (CART) [1] ou les réseaux neuronaux [2] ont été proposées avec un succès limité en recherche clinique ;
- estimation des paramètres et tests éventuels.

Si on note le modèle obtenu $Y=f(X_1,...,X_K| \alpha_1, ..., \alpha_K)$ où f est une fonction de lien, et α' le vecteur des paramètres du modèle, le score de risque individuel est alors la simple application du modèle. Dans le cas de Y binaire, il exprimera la probabilité de développer un événement. Dans le cas d'une analyse de données de survie, il exprimera le délai attendu avant l'apparition de l'événement ou la probabilité qu'il se produise dans un délai donné. On peut faciliter le calcul d'un score pronostique en pratique clinique en divisant chaque α_k par $\min(\alpha_1, ..., \alpha_K)$ et en arrondissant les valeurs obtenues, ce qui conduit généralement à un score de qualité pratiquement équivalent. Il n'aura plus d'interprétation quantitative et sera donc utilisé uniquement pour discriminer les patients entre eux. Dans le cas de données de survie, si un modèle semi-paramétrique est utilisé, aucune prédiction ne pourra être effectuée à moins d'estimer le risque instantané pour les valeurs de références des variables, ce qui requiert de larges échantillons. Un exemple peut être trouvé dans Gail *et al.* [3] pour la prédiction du risque de développer un cancer du sein ou dans les scores pronostiques *Nottingham Prognostic Index* [4] et *Adjuvant Online* [5] pour la prédiction du risque d'une métastase d'un cancer du sein opéré.

Le score pronostique pourra ensuite être découpé en plusieurs valeurs pour former des classes selon des centiles d'intérêt ou à partir de mesures statistiques et correspondre aux options interventionnelles du clinicien. On supposera par la suite qu'un nombre limité de classes de risque a été dérivé du score. Ainsi, pour l'étude de chimio-prévention par Tamoxifene® chez les femmes à risque élevé de développer un cancer du sein [6], un score de Gail supérieur à 1,65 (correspondant au risque d'une femme de 65 ans de développer un cancer du sein sur les 5 années à venir) a-t-il été utilisé comme critère d'inclusion.

Nombre de sujets nécessaires

Comme pour toute construction de modèles multivariés, le nombre de sujets nécessaires pour obtenir une puissance donnée requiert de connaître le nombre de variables et les interactions entre les paramètres, rendant son calcul impossible en pratique. Harrell *et al.* [7] recommandent d'utiliser 10 fois plus d'événements que de variables étudiées. Une série rétrospective d'une centaine de patients peut donc souvent suffire pour un petit nombre de variables. Dans le cas de la

génomique à haut débit, des techniques particulières sont proposées mais avec un succès limité (*cf.* chapitre IV.6 « Analyse statistique des données d'expression de gènes issues de puces à ADN», page 254).

Évaluation de la capacité pronostique

Dans de nombreux articles, la « qualité » d'un score est mesurée par le degré de signification du test d'association *(p-value)* entre le score et l'événement à expliquer. Lorsque différents scores sont comparés, trop d'auteurs se contentent de produire des critères d'information comme l'Akaike (AIC) ou le critère bayésien (BIC) qui combinent vraisembleblance des paramètres et nombre de paramètres. Cela ne saurait être suffisant puisque le *p* permet uniquement de porter un jugement de signification et indique s'il existe une association, ce qui est une information de peu d'intérêt, et la vraisemblance nous renseigne sur la quantité d'information portée par les données modélisées et est donc inutilisable pour mesurer la capacité pronostique. Pourtant une littérature abondante présente les meilleures approches pour l'évaluation de la capacité pronostique [7, 8]. Deux notions doivent être distinguées : la calibration et la discrimination.

• La première est la capacité du modèle à correctement prédire le niveau de risque pour un groupe défini par une combinaison des variables pronostiques. Elle correspond au biais de l'estimation de f(X). Par définition, la calibration du modèle sera bonne sur l'échantillon utilisé pour la construction du modèle. Sur un autre échantillon, le modèle pourra être aisément recalibré sans modifier sa capacité discriminante. Ainsi, le score de Gail, initialement déterminé sur la population américaine, a été récemment calibré sur la population française *via* la cohorte E3N [9].

• La discrimination est la capacité à séparer les différentes classes du score pronostique. Elle résulte de la qualité informative du score, à savoir la quantité d'information apportée par les variables du modèle, de la séparabilité entre les différentes classes du score ainsi que de la capacité prédictive. Cette dernière est étroitement liée aux deux premières. Toute évaluation d'un score devrait présenter ces deux aspects.

La discrimination est souvent exprimée par la concordance. Soit l'ensemble de toutes les paires de patients ; pour chacune, on peut ordonner les valeurs prédites et les valeurs observées. La mesure proposée par Harrell [7] est alors le ratio du nombre de paires concordantes sur l'ensemble des paires possibles. Le C de Harrell fournit une mesure facile à calculer et à interpréter. Une valeur de 0,5 correspond à une absence de concordance et la valeur 1 à une concordance parfaite. Dans le cas où Y est binaire, il correspond à l'aire sous la courbe (AUC pour *Area Under the Curve*) ROC *(Receiver Operating Characteristic)*. Il est également relié au coefficient de corrélation sur les rangs de Spearman et dérive d'une modification du τ de Kendal. Une extension aux données de survie a été proposée [7] : les temps avant événement sont alors comparés aux rangs du score de risque. Un patient de meilleur pronostique devrait faire l'événement plus tardivement. Certaines paires ne peuvent être utilisées en cas de censure avant l'occurrence de l'événement chez le sujet le plus à risque. Si on reprend le modèle de Gail, on constate que si sa calibration est bonne, le c-index n'est que de 0,63 sur la population d'E3N.

Si cette mesure est la plus répandue, elle n'est affectée que par des variables pronostiques très fortement associées à l'événement et fréquentes, ce qui est malheureusement rarement le cas [10]. De nouvelles mesures de discrimination plus sensibles ont donc été recherchées.

Royston *et al.* ont proposé une mesure de séparabilité [11]. Elle correspond au coefficient de régression pour la variable score dans un modèle où est prédit le rang des patients pour leur réalisation de Y. Cette mesure n'est pas standardisée au sens où par construction elle n'admet pas de valeur maximum, mais elle permet de comparer des scores entre eux. Une correction pour le sur-optimisme est proposée lorsqu'on ne dispose pas d'un échantillon indépendant pour la validation. Elle n'est pas très éloignée dans sa philosophie de la comparaison des pentes de régressions proposées par Yates [12]. Plus récemment, Pencina *et al.* [13] ont proposé deux mesures assez sensibles pour détecter des améliorations à des scores pronostiques lorsqu'on ajoute des variables supplémentaires. Le *net reclassification index* (NRI) est développé sur les tables de reclassification. Ces tables présentent les pourcentages de sujets avec événement (respectivement sans événement) qui sont mis dans une catégorie plus à risque (respectivement moins à risque) avec le nouveau score. Pour un sujet *i*, *v(i)* prend la valeur 1 ou moins 1 si le mouvement est vers une catégorie plus élevée ou moins élevée. Le NRI quantifie ces mouvements :

$$\frac{\sum_{\text{événements } i} v(i)}{\#\ \text{événements}} - \frac{\sum_{\text{non événements } j} v(j)}{\#\ \text{non événements}}$$

Si on considère maintenant un score pronostique continu plutôt que catégoriel, les auteurs proposent de quantifier les mouvements des probabilités prédites. Si cette probabilité augmente, *v(i)* prend la valeur 1, pondérée par la différence entre les valeurs prédites avec le nouveau score et les valeurs prédites avec l'ancien.

La seconde mesure, dite *integrated discrumination improvement* (IDI), est alors la différence :

$$\frac{\sum_{\text{événements } i} (\hat{p}_{new}(i) - \hat{p}_{old}(i))}{\#\ \text{événements}} - \frac{\sum_{\text{non événements } j} (\hat{p}_{new}(j) - \hat{p}_{old}(j))}{\#\ \text{non événements}}$$

Le premier terme mesure l'amélioration de la sensibilité, tandis que la valeur négative du second quantifie l'amélioration de la spécificité.

Pour chaque cas, un test d'égalité à 0 est proposé dans le cas où les deux scores sont calculés sur le même échantillon. Comme ils reposent sur des séries appariées, ils sont assez puissants. Un exemple sur les équations de Framingham montre que la valeur ajoutée des lipoprotéines à densité haute pour prédire un risque de maladie cardiovasculaire est significative alors que la différence entre les deux AUC était non significative [13]. Ils montrent en outre les connexions entre l'IDI et l'indice de Youden, la différence des pentes de corrélation de Yates et son interprétation en termes de sensibilité moyenne. Le choix de la meilleure mesure dépend du contexte : le NRI est préférable si des classes existent. Des extensions aux données de survie sont en cours de développement [14].

Enfin, la capacité prédictive d'un modèle est la mesure qui évalue la distance entre les prédictions du modèle et les observations. Elle dépend donc largement du choix du modèle. Dans le cas d'un modèle linéaire, voire linéaire généralisé, l'écart moyen attendu qui représente le carré du biais plus la variance de l'estimation est une mesure naturelle de la qualité de la prédiction. Elle correspond au R^2. En revanche, dans le cas des données de survie, son usage dépend du taux de données censurées la rendant difficilement interprétable. L'estimation de la proportion expliquée a également été étendue au cas des données de survie avec différentes formulations, par exemple dans [15-17]. Elles partagent en général une dépendance plus ou moins forte au taux de données censurées.

Validation de la capacité pronostique

L'examen de la capacité pronostique ou discriminante *apparente* d'un modèle multivarié sur l'échantillon qui sert à construire le modèle (ou échantillon d'apprentissage) est d'une utilité limitée. En effet, mesurer la capacité pronostique sur l'échantillon d'apprentissage en ignorant les fluctuations d'échantillonnage conduit à un « sur-optimisme » qui se traduit par une surestimation des qualités du modèle. Une validation externe (sur un autre échantillon que celui ayant généré le modèle) devrait toujours être réalisée. Lors de la présentation des résultats, seules les valeurs obtenues sur l'échantillon de validation devraient être fournies. En l'absence d'échantillon de validation, certaines techniques permettent de corriger ce sur-optimisme.

Découpage de l'échantillon

La plus répandue consiste à couper l'échantillon en deux parties. La première partie sert à l'apprentissage, le modèle est ensuite gelé et appliqué à la seconde partie (échantillon de validation) pour calculer les mesures de discrimination et de qualité prédictive. Le découpage peut être aléatoire mais cette procédure ignore alors une partie de la variabilité [18, 19] et on lui préférera un sous-échantillonnage non aléatoire, par exemple par période ou par centre. La taille de chaque sous-échantillon doit permettre une estimation du modèle sans trop de perte de puissance et une estimation de la qualité de la prédiction. Il semble donc préférable d'avoir un échantillon plus grand pour l'apprentissage. Une proportion de 2/3 est parfois suggérée [20]. Cette approche, même si elle corrige une partie du biais de surestimation des performances, induit toutefois une perte non négligeable de puissance [21]. Les méthodes de *cross-validation* par *bootstrap* ou *leave-one-out* offrent de meilleurs résultats [21].

La validation croisée

La *cross-validation* consiste à répéter le découpage de l'échantillon. À chaque répétition, il est essentiel de reconstruire complètement le modèle en utilisant systématiquement la même procédure, de réestimer les paramètres et d'appliquer le résultat sur l'échantillon de validation. Par exemple, sur un échantillon de taille n, le modèle peut être construit sur un sous-échantillon aléatoire correspondant à 90 % des données et être appliqué sur les 10 % restant. Cette étape est répétée 10 fois. À l'extrême, toutes les observations sauf une servent à l'apprentissage. Et cette étape est répétée *n* fois. Le bénéfice sur la méthode précédente est clair puisqu'elle utilise une

taille d'échantillon pour l'apprentissage bien plus grande. Toutefois, Efron [22] a montré que la validation croisée restait assez inefficace du fait de la grande variabilité des mesures de validation obtenues.

Le *bootstrap*

Cette technique est définie par le ré-échantillonnage avec remise [21]. Les sous-échantillons sont alors de taille *n*, comme l'échantillon initial, assurant d'utiliser la totalité de l'information tant pour l'apprentissage que pour la validation [22]. Comme pour l'approche précédente, il faut que toutes les étapes de la construction du modèle soient répétées. Cela suppose donc qu'une procédure automatisée soit utilisée.

REMARK

Devant le nombre important d'études pronostiques insuffisamment renseignées dans la littérature médicale, des recommandations pour la rédaction de telles communications ont été proposées : *REporting recommendations for tumour MARKer prognostic studies* (REMARK) [23]. Elles sont présentées dans le *tableau I*.

Tableau I. Recommandations pour la publication des études pronostiques des marqueurs tumoraux (REMARK).
Introduction 1. Affirmer quel marqueur sera étudié, préciser les objectifs de l'étude et préspécifier les hypothèses
Matériel et méthodes ***Patients*** 2. Décrire les caractéristiques (*e.g.* stade de la maladie ou comorbidités) des patients à l'étude, en précisant leur provenance ainsi que les critères d'inclusion et exclusion 3. Décrire les traitements reçus et comment ils ont été choisis (*e.g.* randomisé ou protocolaire)
Caractéristiques des spécimens 4. Décrire le type de matériel biologique utilisé (comprenant les échantillons témoin), ainsi que les méthodes de conservation et stockage
Méthode d'essai 5. Spécifier la méthode d'essai utilisée et fournir (ou référence à) un protocole détaillé, comprenant les réactifs spécifiques ou des kits utilisés, les procédures de contrôle qualité, les évaluations de reproductibilité, les méthodes de quantification, ainsi que le *scoring* et les protocoles de rapports. Spécifier si et comment les essais ont été réalisés en aveugle du critère principal
Plan expérimental 6. Affirmer quelle méthode a été utilisée pour sélectionner les cas, en indiquant s'il s'agit d'une sélection prospective ou rétrospective et si une stratification ou un appariement (*e.g.* par stade de la maladie ou âge) a été utilisé. Spécifier la période de temps où les cas ont été sélectionnés, la fin de la période de suivi, ainsi que le suivi médian 7. Définir de manière précise tous les critères d'évaluation cliniques considérés

8. Lister toutes les variables candidates qui ont été étudiées initialement ou considérées pour l'inclusion dans les modèles
9. Donner le rationnel pour le calcul de la taille de l'échantillon ; si l'étude a été planifiée pour détecter un effet spécifique, donner la puissance ciblée ainsi que la taille de l'effet à détecter

Méthodes d'analyses statistiques
10. Spécifier toutes les méthodes statistiques, comprenant les détails des procédures pour la sélection des variables (si pertinent) ainsi que d'autres considérations prises en compte pour la construction de modèles, comment les hypothèses sous-jacentes aux modèles ont été vérifiées et comment les données manquantes ont été gérées
11. Clarifier comment les valeurs des marqueurs ont été gérées dans les analyses ; si pertinent, décrire les méthodes utilisées pour déterminer les valeurs seuils

Résultats
Données
12. Décrire le parcours des patients pendant l'étude, comprenant le nombre de patients inclus à chaque étape de l'analyse – un diagramme peut être utile – et donner les raisons de sortie d'étude. Plus spécifiquement, de manière globale et pour chaque sous-groupe étudié en détail, rapporter le nombre de patients ainsi que le nombre d'événements
13. Présenter les distributions des caractéristiques démographiques classiques (au moins âge et sexe), les variables pronostiques habituelles (spécifiques de la maladie, ainsi que le marqueur tumoral), y compris le nombre de valeurs manquantes

Analyses et présentation
14. Montrer la relation entre le marqueur et les variables pronostiques habituelles
15. Présenter les analyses univariées en montrant la relation entre le marqueur et le critère principal, avec une estimation de l'effet (*e.g. hazard ratio* et probabilité de survie). De préférence, fournir des analyses similaires pour toutes les autres variables analysées. En ce qui concerne l'effet d'un marqueur tumoral sur un critère de survie, une courbe de Kaplan-Meier est recommandée
16. En ce qui concerne les analyses multivariées clés, donner les estimations des effets (*e.g. hazard ratio*) avec un intervalle de confiance pour le marqueur et, au moins pour le modèle final, toutes les autres variables dans le modèle
17. Parmi les résultats rapportés, fournir les estimations des effets avec les intervalles de confiance à partir d'une analyse dans laquelle le marqueur et les variables pronostiques classiques ont été inclus, quel que soit le degré de leur signification
18. S'ils ont été faits, rapporter les résultats des autres investigations, telles que la vérification des hypothèses, analyses de sensibilité, validation interne

Discussion
19. Interpréter les résultats dans le contexte des hypothèses préspécifiées, ainsi que toute autre étude pertinente ; inclure une discussion des limites de l'étude
20. Discuter les implications pour la recherche future ainsi que la valeur clinique

Facteurs prédictifs de la réponse à un traitement

Tests d'interaction

Un facteur (ou marqueur) prédictif prédit la réponse à un traitement. On considère ici qu'il peut prendre deux valeurs : positive ou négative, même s'il résulte d'une combinaison des variables. Deux approches sont communément utilisées pour identifier un facteur prédictif : le test d'interaction et l'étude en sous-groupe. En théorie, seul le test d'interaction permet de conclure à un effet du traitement différentiel selon la valeur du marqueur. Pour rappel, on distingue l'interaction quantitative, où l'effet du traitement est de grandeur différente selon le marqueur mais va dans le même sens, et l'interaction qualitative [24], où l'effet du traitement est opposé (bénéfique et délétère). Par extension, la situation d'un effet dans un groupe et pas d'effet dans l'autre est également incluse dans ce groupe. En pratique, l'interaction quantitative est attendue à des degrés divers dans de nombreuses situations et présente peu d'intérêt. En général, l'effet du traitement peut alors être résumé par l'effet moyen sur les différents groupes.

Plans d'expérience pour la validation

Sargent *et al.* [25] ont présenté divers plans d'expérience pour valider un facteur prédictif de la réponse à un traitement en distinguant l'étude directe de l'étude indirecte. Nous supposerons que les informations préliminaires motivant l'existence d'un tel marqueur ont déjà été obtenues d'une étude rétrospective ou d'une étude de cohorte antérieure.

Étude indirecte *(figure 2)*

Il s'agit de deux plans d'expérience assez similaires sauf pour leur analyse. Le principe de ces plans dit d'interaction est de randomiser l'attribution des deux traitements dans chaque strate du marqueur. Ainsi, l'effet relatif des deux traitements peut-il être mesuré pour chaque valeur du marqueur. Un test formel d'interaction ou des tests dans chacun des sous-groupes peuvent être réalisés. Un exemple ambitieux est fourni par l'étude 10994 du groupe « sein » de l'*European organization for research and treatment of cancer* (EORTC) qui étudie prospectivement la valeur prédictive des mutations *P53* sur la chimio-sensibilité aux traitements néo-adjuvants à base de taxanes ou d'anthracyclines [26] : 1 440 patientes ont été randomisées entre un bras sans et un bras avec taxane après obtention de prélèvements afin de tester, au seuil global de 5 %, la différence globale entre les traitements et également la différence dans chacun des sous-groupes (*P53* mutée et non mutée).

Étude directe *(figure 3)*

Elle se décline en deux autres plans où la randomisation détermine si le traitement des patients sera guidé ou pas par la valeur du marqueur. Supposons que le nouveau traitement B soit à l'étude.

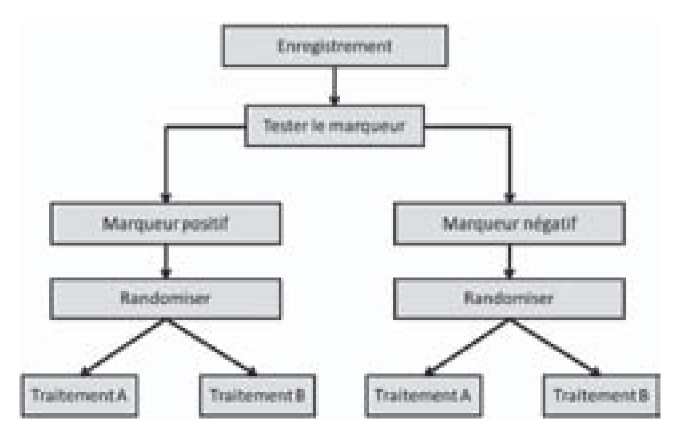


Figure 2. Plan d'expérience pour une étude indirecte du caractère prédictif.

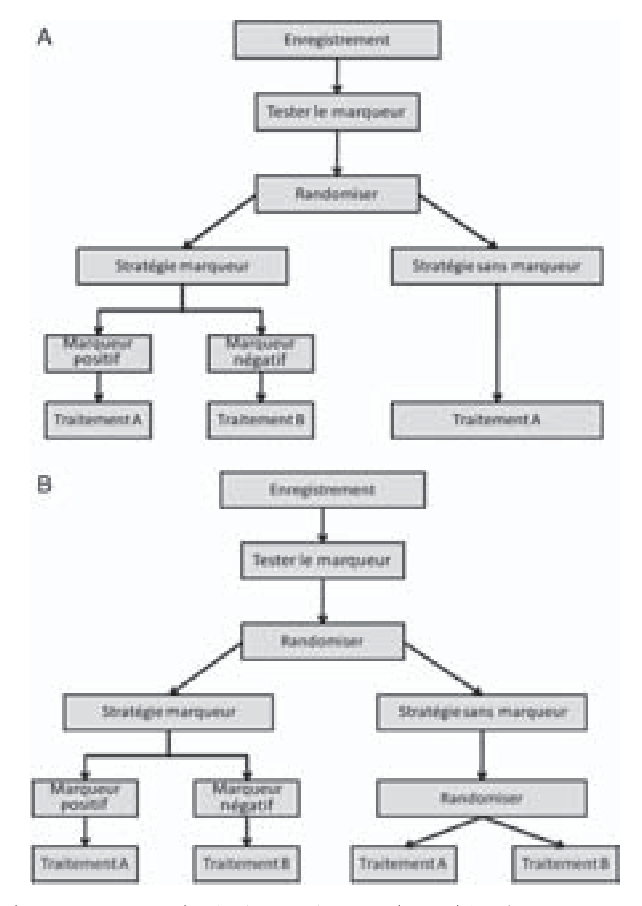


Figure 3. Plan d'expérience pour une étude directe du caractère prédictif.
A. *Sans* randomisation du traitement dans le bras non guidé par le marqueur.
B. *Avec* randomisation du traitement dans le bras non guidé par le marqueur.

La moitié des patients recevra le traitement standard A indépendamment de la valeur de leur marqueur et l'autre moitié recevra A ou B selon le marqueur. On compare alors les deux bras randomisés. Un des inconvénients est que la comparaison des deux traitements se fait sur des effectifs fortement déséquilibrés. Une modification de ce plan consiste en une seconde randomisation pour attribuer aléatoirement A ou B dans le bras non guidé par le marqueur. Cette seconde randomisation permet de distinguer un effet prédictif de la réponse au traitement de l'effet du traitement B comparé à A. De tels plans devraient prendre en compte un possible déséquilibre dans la prévalence du marqueur.

Les auteurs montrent par des simulations que les schémas indirects requièrent en fait relativement moins de patients que les schémas directs. D'autre part, le test d'une interaction requiert moins d'événements que la mise en évidence d'effets du traitement dans chacun des groupes, lorsque l'effet de l'interaction est du même ordre que l'effet du traitement, ce qui est généralement le cas en cancérologie. Le protocole de l'essai de phase III du panitumumab *versus* soins palliatifs dans le cancer colorectal métastatique avait ainsi prévu de tester l'interaction entre le statut KRAS et la survie sans progression [27]. L'analyse de 427 patients a permis de montrer qu'il existait un effet différentiel selon le marqueur (test d'interaction $p < 0{,}001$) : relativement fort (HR = 0,45) dans le groupe KRAS sauvage et non significatif (HR = 0,99) lorsque KRAS était muté. Le test sur l'ensemble de la population était significatif (HR = 0,56).

Calcul du nombre de sujets nécessaires

Nous avons calculé les tailles d'échantillons pour mettre en évidence une interaction suivant un schéma indirect sur un critère binaire ou de survie *(figure 2)*. Une solution dérivée de l'analyse de variance utilisant une distribution F non centrale existe dans les logiciels Nquery et SAS dans le cas d'un tableau 2x2. Pour les données de survie, on peut se baser sur les calculs des tailles d'échantillons d'un plan 2x2 factoriel [28] Ces différents calculs reposent sur un modèle multiplicatif qui correspond aux interactions pour les *odds ratios* ainsi qu'à l'introduction d'un terme d'interaction dans le modèle Cox.

La *figure 4* montre un exemple du nombre de sujets nécessaires selon l'amplitude de l'effet à mettre en évidence. On se place dans la situation favorable où la prévalence du marqueur est de 50 %. On suppose un taux d'événements dans le groupe contrôle de 40 % et 50 % selon que les sujets sont M+ ou M- (facteur pronostique). Pour mettre en évidence un rapport entre les deux effets du traitement chez les M+ et les M- d'un OR = 0,5, il faut plus de 1 000 patients. Ces effectifs deviennent vite irréalistes et on retrouve la notion habituelle qu'un test d'interaction est peu puissant si on ne dispose pas de centaines dévénements.

C'est pourquoi, il est fréquent de voir des plans d'expérience où l'analyse en sous-groupe n'est qu'un objectif secondaire avec une puissance plus faible. L'objectif principal est alors de tester l'effet sur l'ensemble de la population et, si le test n'est pas rejeté, de tester un sous-groupe d'intérêt préalablement identifié. La question est alors de choisir comment dépenser le risque de première espèce pour maintenir un risque global de 5 %. Simon *et al.* [29], entre autres, ont

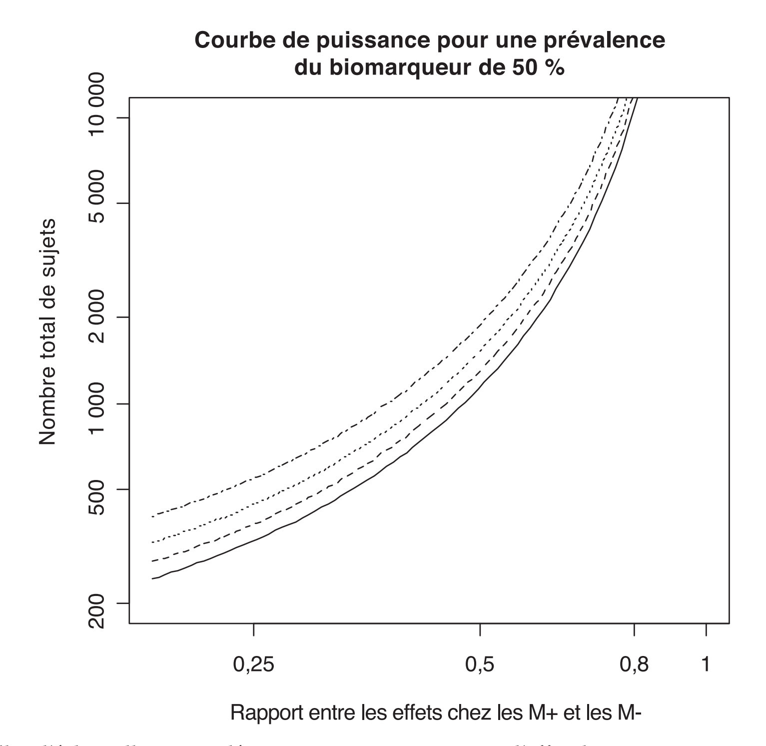


Figure 4. Tailles d'échantillon pour détecter une interaction entre l'effet du traitement et un biomarqueur de prévalence 50 % avec une puissance de 80 %.

suggéré de faire le test principal à 4 % et le test dans le sous-groupe à 1 %. Une telle procédure ne fournit pas le même niveau de preuve qu'un test d'interaction mais permet de contrôler la taille de l'échantillon.

Conclusions

Mettre en évidence un facteur pronostique qui sera ensuite appliqué en clinique résulte d'une démarche rigoureuse et objective pour correctement quantifier son apport. Le facteur pronostique doit impérativement être validé, si possible sur une série indépendante. La construction et la quantification de la valeur pronostique est facilement réalisable, notamment sur des séries rétrospectives. La démonstration de l'utilité en clinique pose d'autres questions qui n'ont pas été abordées dans ce chapitre. On pourra se référer aux discussions entourant l'essai MINDACT pour creuser ce point [30]. Il est utile de garder en mémoire qu'une démonstration rigoureuse de l'utilité d'un facteur pronostique requiert de grands effectifs. La mise en évidence d'un facteur prédictif de la réponse à un traitement donné requiert de larges effectifs et des plans d'expérience prospectifs. Bien que la recherche de marqueurs biologiques associés à la réponse à des traitements soit florissante dans la littérature médicale, bien peu d'études sont construites pour les

valider et leur application clinique reste encore assez exceptionnelle. Il est très probable qu'un certain nombre de facteurs décrits comme prédictifs de la réponse à un traitement suite à des plans d'expériences inadéquats (études de phase II non randomisées notamment) ne soient que des facteurs pronostiques.

Références

1. Austin PC. A comparison of regression trees, logistic regression, generalized additive models, and multivariate adaptive regression splines for predicting AMI mortality. *Stat Med* 2007 ; 26 (15) : 2937-57.

2. Faraggi D, LeBlanc M, Crowley J Understanding neural networks using regression trees: An application to multiple myeloma survival data. *Stat Med* 2001 ; 20 (19) : 2965-76.

3. Gail MH, Brinton LA, Byar DP, *et al*. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989 ; 81 (24) : 1879-86.

4. Galea MH, Blamey RW, Elston CE, Ellis IO. The Nottingham Prognostic Index in primary breast cancer. *Breast Cancer Res Treat* 1992 ; 22 (3) : 207-19.

5. Ravdin PM, Siminoff LA, Davis GJ, *et al*. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. *JCO* 2001 ; 19 (4) : 980-91.

6. Gail MH. Personalized estimates of breast cancer risk in clinical practice and public health. *Stat Med* 2011 ; 30 (10) : 1090-104.

7. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996 ; 15 (4) : 361-87.

8. Chatfield C. Model uncertainty, data mining and statistical inference *(with discussion)*. *Journal of the Royal Statistical Society*, Series A 1995 ; 158 : 419-66.

9. Viallon V, Ragusa S, Clavel-Chapelon F, Bénichou J. How to evaluate the calibration of a disease risk prediction tool. *Stat Med* 2009 ; 28 (6) : 901-16.

10. Pepe MS, Janes H, Longton G, Leisenring W, Newcomb P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004 ; 159 (9) : 882-90.

11. Royston P, Sauerbrei W. A new measure of prognostic separation in survival data. *Stat Med* 2004 ; 23 (5) : 723-48.

12. Yates JF. External correspondence: Decomposition of the mean probability score. *Organizational Behavior and Human Performance* 1982 ; 30 : 132-56.

13. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med* 2008 ; 27 (2) : 157-72 ; discussion 207-12.

14. Pencina MJ, D'Agostino RB Sr, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med* 2011 ; 30 (1) : 11-21.

15. Kent JT, O'Quigley J. Measures of dependence for censored survival data. *Biometrika* 1988 ; 75 (3) : 525-34.

16. Korn LK, Simon R. Explained residual variation, explained risk and goodness of fit. *American Statistician* 1991 ; 45 (3) : 201-6.

17. Schemper M. The explained variation in proportional hazards regression. *Biometrika* 1990 ; 77 (l) : 216-8 (correction: *Biometrika* 1994 ; 81 (3) : 631).

18. Feinstein AR. *Multivariable Analysis: an Introduction.* New Haven: Yale University Press, 1996 : 184-7, 578-82.

19. Hirsch RP. Validation samples. *Biometrics* 1991 ; 47 : 1193-4.

20. Cox DR. A note on data-splitting for the evaluation of significance levels. *Biometrika* 1975 ; 62 : 441-4.

21. Efron B, Tibshirani R. *An Introduction to the Bootstrap.* London: Chapman & Hall, 1993 : 255.

22. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* 1989 ; 78, 316-31.

23. McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM; Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies (REMARK). *J Natl Cancer Inst* 2005 ; 97 (16) : 1180-4.

24. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985 ; 41 (2) : 361-72.

25. Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005 ; 23 (9) : 2020-7.

26. Therasse P, Carbonnelle S, Bogaerts J. Clinical trials design and treatment tailoring: General principles applied to breast cancer research. *Crit Rev Oncol Hematol* 2006 ; 59 (2) : 98-105.

27. Amado RG, Wolf M, Peeters M, *et al.* Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol* 2008 ; 26 (10) : 1626-34.

28. Peterson B, George SL. Sample-size requirements and length of study for testing interaction in a 2 x k factorial design when time-to-failure is the outcome. *Controlled Clinical Trials* 1993 ; 14 : 511-22.

29. Simon R. Clinical trials for predictive medicine: New challenges and paradigms. *Clin Trials* 2010 ; 7 (5) : 516-24.

30. Buyse M, Sargent DJ, Grothey A, Matheson A, de Gramont A. Biomarkers and surrogate end points-the challenge of statistical validation. *Nat Rev Clin Oncol* 2010 ; 7 (6) : 309-17.

Événements à risque compétitif

C. Bellera, T. Filleron

En cancérologie, en particulier dans l'évaluation de l'efficacité de traitements ou dans le cadre d'études pronostiques, le délai d'apparition d'un événement clinique est un critère de jugement couramment utilisé. L'objectif est alors d'estimer la probabilité que cet événement survienne au-delà d'un certain temps t.

Dans certaines circonstances, on s'intéresse à l'apparition au cours du temps d'un événement clinique pertinent comme la rechute locale, mais l'analyse peut-être compliquée par la présence d'événements à risque compétitif *(competing risks)* ou événements concurrents [1], qui peuvent se produire avant l'apparition de l'événement d'intérêt et ainsi modifier la probabilité d'observer ce dernier, voire empêcher sa réalisation. À titre d'exemple, considérons l'évaluation de l'efficacité d'un traitement dit locorégional, comme la radiothérapie ou la chirurgie. Un critère couramment utilisé dans le cadre d'essais cliniques de phase II en chirurgie ou radiothérapie est le délai d'apparition de la rechute locorégionale, couramment appelé survie sans rechute locale (bien que le terme « survie » ne fasse pas référence au statut vital, mais soit ici synonyme d'« indemne de »). Dans ce contexte, l'apparition de métastases avant la rechute locorégionale est considérée comme un événement à risque compétitif. En effet, on considère que cette probabilité est affectée par la survenue de métastases.

Une fois que les événements à risque compétitif sont identifiés, se pose alors la question de leur prise en compte dans l'analyse statistique des données. Dans l'exemple, l'événement d'intérêt est le délai jusqu'à la rechute locale. Comment devons-nous analyser les données des sujets dont le premier événement observé est l'apparition de métastases ? Dans la littérature médicale, en présence d'un événement compétitif, le délai jusqu'à l'événement principal est analysé selon différentes approches :

- seul l'événement d'intérêt est pris en compte et les autres types d'événements survenus sont ignorés ;
- seul le premier événement est pris en compte, qu'il s'agisse ou non de l'événement d'intérêt ; les données sont censurées à la date d'observation de ce premier événement s'il ne correspond pas à l'événement d'intérêt.

Ces deux approches sont respectivement appelées méthodes *Ignore* et *Censure* [2]. Ces approches ne sont cependant pas optimales et peuvent conduire à des estimations biaisées. Une troisième méthode, nommée *Inclure*, permet de prendre en considération les événements compétitifs en les modélisant explicitement et ainsi d'obtenir des estimations non biaisées.

Ces trois approches sont détaillées ainsi que leurs conséquences éventuelles sur les estimations. Nous présentons tout d'abord la terminologie relative aux risques compétitifs ainsi qu'un exemple qui nous permet d'illustrer les différentes méthodes exposées.

Définitions

Une cohorte de patients est suivie pendant une période d'observation définie par le temps écoulé entre une date d'origine (par exemple, la date de diagnostic, de début du traitement ou de randomisation) jusqu'à une date de point ou de dernières nouvelles. *K* types d'événements différents peuvent être observés durant cette période.

Survie sans événement

La survie sans événement à un instant *t*, *SSE(t)*, est définie par le délai entre la date d'origine et la date d'apparition du premier des K événements ou, si aucun événement n'est observé, par le délai jusqu'à la date de dernières nouvelles (les données sont alors censurées). La *SSE(t)* est estimée à chaque instant *t* et représente la probabilité d'être indemne de tous les événements considérés à l'instant *t*. Cette probabilité peut être estimée par la méthode d'estimation non paramétrique de Kaplan-Meier (*cf.* chapitre III.1 « Données de survie », page 129).

Survie spécifique ou marginale

La survie spécifique de l'événement *k*, $S_k(t)$, $k = 1, \ldots, K$, ou survie marginale, correspond à la probabilité d'être indemne de l'événement *k* à l'instant *t*.

Incidence cumulée spécifique

La fonction d'incidence cumulée de l'événement *k*, $I_k(t)$, représente la probabilité que l'événement *k* se produise avant l'instant *t*. Elle correspond au complémentaire de la fonction de survie spécifique associée : $I_k(t) = 1 - S_k(t)$.

Fonction de risque spécifique

La fonction de risque spécifique $h_k(t)$ pour un événement *k* à un instant *t* correspond à la probabilité (ou risque) instantané de survenue de cet événement à l'instant *t*.

Événements compétitifs indépendants ou exclusifs

Des événements sont dits compétitifs (ou concurrents) lorsque la survenue affecte la probabilité d'observer le second. Ces événements compétitifs peuvent être dépendants ou indépendants.

Deux événements compétitifs sont considérés indépendants lorsque la réalisation de l'un n'affecte pas la probabilité de réalisation de l'autre. Supposons que l'on s'intéresse au risque de développer une complication rénale post-chirurgie. On peut ainsi considérer que le risque de développer une telle complication n'est pas affecté par la survenue de complications infectieuses (c'est ici une hypothèse). Ces deux événements, complication rénale et complications infectieuses, sont donc indépendants. Il est important de noter ici que l'hypothèse d'indépendance doit être justifiée cliniquement.

Dans le cas d'événements dépendants, ceux-ci sont exclusifs ou non exclusifs. On parle d'événements exclusifs dans le cas où la réalisation du premier type d'événement empêche définitivement la réalisation du second. Un exemple classique est l'étude du décès lié au cancer. Si l'on s'intéresse à la probabilité de décéder du cancer, un décès d'une cause autre que le cancer est un événement compétitif puisqu'il affecte la probabilité ultérieure de décéder du cancer. Ces deux événements sont exclusifs : la réalisation de l'un empêche la réalisation du second. Enfin, des événements sont non exclusifs (ou compatibles) lorsque la réalisation de l'un affecte, sans l'annuler, le risque d'observer le second. Dans l'exemple précédent, la rechute locorégionale et l'apparition de métastases sont deux événements non exclusifs.

Exemple

Considérons une cohorte hypothétique de 5 sujets qui sont suivis à partir d'une date d'origine correspondant, par exemple, à leur date de randomisation dans un essai clinique. La durée maximale du suivi est de 12 mois. Durant cette période, il est possible d'observer jusqu'à $K = 2$ événements par sujet : la rechute locorégionale (L) et/ou l'apparition de métastases (M). Ces événements sont non exclusifs puisque tous deux peuvent survenir pour un même patient. De plus, on peut supposer qu'ils sont dépendants, puisque le risque de métastases est différent selon qu'une rechute locorégionale ait été observée antérieurement ou non. L'histoire de vie de ces 5 sujets est représentée de manière verticale dans la *figure 1* où chaque ligne représente le suivi d'un sujet différent (S1 à S5) :

- le 1er sujet, S1, a été suivi 12 mois sans présenter d'événement ;
- pour le 2e sujet, S2, une rechute locorégionale a été observée à 5 mois. Il a ensuite été perdu de vue à 9 mois ;
- le 3e sujet, S3, a développé des métastases à 3 mois, puis a été perdu de vue à 9 mois ;
- le 4e sujet, S4, a rechuté loco-régionalement à 4 mois, développé des métastases à 6 mois, puis a été perdu de vue à 7 mois ;
- le dernier sujet, S5, a développé des métastases à 2 mois, rechuté loco-régionalement à 8 mois et a ensuite été suivi jusqu'à 12 mois.

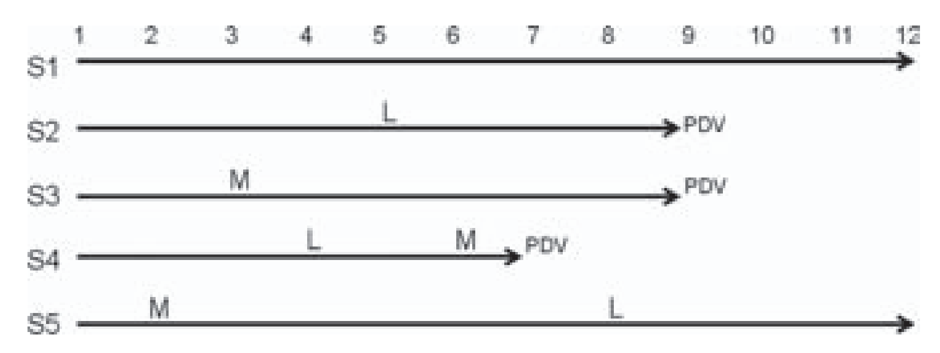


Figure 1. Histoire de vie de 5 sujets.
M : métastase ; L : rechute locorégionale ; PDV : perdu de vue.

On souhaite estimer la survie sans événement *SSE(t)*. On s'intéresse donc à la date du premier événement, c'est-à-dire l'apparition d'une rechute locorégionale ou de métastases. Le calcul du risque instantané *h(t)* et de *SSE(t)* est présenté pour ces 5 patients dans le *tableau I* suivant la méthodologie décrite dans le chapitre III.1 (*cf.* page 129). Les sujets sont tout d'abord ordonnés en fonction du temps d'observation du premier événement. Dans notre exemple, une rechute locorégionale ou des métastases ont été observées pour 4 patients, conduisant à une estimation de la survie sans événement à 12 mois de 20 %.

L'estimation de la survie sans événement ne présente aucune ambiguïté. En revanche, l'estimation des incidences cumulées et survies spécifiques pour chaque type d'événement, rechute locorégionale ou survenue de métastases, peut être appréhendée selon trois approches que nous détaillons ci-dessous ainsi que leurs hypothèses sous-jacentes.

Tableau I. Survie sans événement (ni rechute locorégionale, ni métastase).

Temps (t)	Sujets	Événement	À risque	*h(t)*	1 - *h(t)*	*SSE(t)*
0		0	5	0	0	1
2	S5	1	5	1/5	4/5	4/5 = 80 %
3	S3	1	4	1/4	3/4	3/5 = 60 %
4	S4	1	3	1/3	2/3	2/5 = 40 %
5	S2	1	2	1/2	1/2	1/5 = 20 %
12	S1	0	1	0	0	1/5 = 20 %

Méthode *Ignore*

Principe général

Une première approche d'analyse des risques compétitifs consiste à définir un événement unique qui sera considéré comme événement principal lors de l'analyse des données. Les autres événements sont totalement ignorés, qu'ils soient observés avant ou après l'événement principal.

Définition des délais et des événements

Lorsque l'événement principal est observé pour un sujet donné, le temps de participation (TP) est défini par le délai jusqu'à la date de son apparition. Concernant les sujets pour lesquels l'événement principal n'a pas été observé, leurs TP sont censurées à la date de fin d'étude (dernières nouvelles), qu'un événement compétitif ait été observé ou non.

Estimation de la survie et de l'incidence cumulée spécifiques

Une fois les temps de participation et d'événements établis pour chaque sujet, la survie spécifique associée à l'événement *k*, $S_k^{Ign}(t)$, c'est-à-dire la probabilité d'être indemne de l'événement *k* au temps *t*, peut donc être estimée à partir de la méthode de Kaplan-Meier.

Avec cette méthode, on suppose que le risque de développer l'événement principal n'est pas affecté par l'apparition d'un événement compétitif. Cette hypothèse n'est vraie qu'en présence d'événements compétitifs indépendants et les estimations de survie et d'incidence cumulée spécifiques ne seront donc pas biaisées si cette hypothèse est respectée. À défaut, en présence d'événements dépendants, exclusifs ou non exclusifs, les estimations seront biaisées : l'incidence cumulée spécifique sera surestimée.

Estimation de la survie sans événement

La survie sans événement est estimée par la méthode de Kaplan-Meier en considérant le délai d'apparition du premier des *K* événements quel qu'il soit, que les événements soient exclusifs ou non. Dans le cas particulier d'événements indépendants, la survie sans événement correspond au produit des survies spécifiques :

$$SSE(t) = S_1^{Ign}(t) \times S_2^{Ign}(t) \times \ldots \times S_K^{Ign}(t)$$

Exemple fil rouge

Les données de survie pour un sujet *i* sont résumées par la paire de variables (t_i, c_i), $t_i \geq 0$ et $c_i = 0$ ou 1, où t_i correspond au temps jusqu'à l'événement d'intérêt, et c_i est une variable indicatrice égale à 1 si l'événement est observé ou 0 si les données sont censurées. Dans notre exemple, la

variable t_i correspond donc au délai jusqu'à la date de rechute locorégionale ou la date de dernières nouvelles. Avec la méthode *Ignore*, les données des 5 patients hypothétiques de la *figure 1* sont représentées de la manière suivante :

• sujet S1 : aucun événement n'a été observé, ses données sont donc censurées à la date de dernières nouvelles et $(t_1, c_1) = (12,0)$;
• sujet S2 : la rechute locorégionale est observée à 5 mois, ainsi $(t_2, c_2) = (5 ; 1)$;
• sujet S3 : aucune rechute locorégionale n'est observée et le suivi s'est achevé à 9 mois. L'apparition de métastases n'étant pas prise en compte dans la méthode *Ignore*, on notera donc $(t_3, c_3) = (9, 0)$;
• sujet S4 : la rechute locorégionale est observée à 4 mois, ainsi $(t_4, c_4) = (4 ; 1)$;
• sujet S5 : la rechute locorégionale est observée à 8 mois. L'apparition de métastases à 2 mois étant ignoré, on a donc $(t_5, c_5) = (8 ; 1)$.

Concernant la rechute locorégionale, lorsque l'approche *Ignore* est appliquée, des rechutes locorégionales sont observées aux temps 4, 5 et 8. Les fonctions de risque instantané h_L^{Ign}, d'incidence cumulée I_L^{ign} et de survie spécifique à la rechute locorégionale S_L^{Ign} peuvent être calculées à chacun de ces temps à partir de la méthode de Kaplan-Meier *(tableau II)*. À 12 mois, la survie sans rechute locorégionale est estimée à 40 % lorsque l'on ignore les événements compétitifs.

De la même manière, le délai d'apparition des métastases est calculé après avoir préalablement redéfini les variables (t_i, c_i) où, cette fois-ci, t_i correspond au délai jusqu'à l'apparition de métastases et c_i est la variable indicatrice égale à 1 si cet événement est observé ou 0 en cas de données censurées. À 6 mois, la survie sans métastase est estimée à 40 % lorsque l'on ignore les événements compétitifs *(tableau II bis)*.

Tableau II. Application de la méthode *Ignore* (rechutes locorégionales).

T (sujet)	Événement	À risque	$h_L^{Ign}(t)$	$1 - h_L^{Ign}(t)$	$S_L^{Ign}(t)$
0	–	–	–	1	1
4 (S4)	1 (S4)	5	1/5	4/5	4/5 = 80 %
5 (S2)	1 (S2)	4	1/4	3/4	3/5 = 60 %
8 (S5)	1 (S5)	3	1/3	2/3	2/5 = 40 %

Pour un même sujet, il est cependant possible d'observer une rechute locorégionale ainsi que des métastases ; ces deux événements ne sont donc pas exclusifs. Dans ce contexte, la probabilité d'observer une rechute ou une métastase ne peut pas être calculée à partir de la somme des deux fonctions d'incidence cumulée obtenues par la méthode *Ignore* : $I_{LM}^{ign}(t) \neq I_L^{ign}(t) + x\, I_M^{ign}(t)$. De même, ces deux événements n'étant pas des événements indépendants, le produit de leur probabilité de survie spécifique ne correspond pas à la survie sans événement $SSE(t) \neq S_L^{ign}(t)\ x\ S_M^{ign}(t)$. L'incidence cumulée et la survie sans événement doivent donc être calculées autrement afin d'obtenir des estimations non biaisées, c'est-à-dire en considérant le délai jusqu'au premier de ces événements.

Tableau II bis. Application de la méthode *Ignore* (métastases).

T (sujet)	Événement	À risque	$h_M^{Ign}(t)$	$1 - h_M^{Ign}(t)$	$S_M^{Ign}(t)$
0	–	–	–	1	1
2 (S5)	1	5	1/5	4/5	4/5
3 (S3)	1	4	1/4	3/4	3/5
6 (S4)	1	3	1/3	2/3	2/5

Avec la méthode *Ignore*

- Seul l'événement d'intérêt principal est pris en compte, qu'un événement compétitif ait été observé antérieurement ou non. Tout sujet est considéré à risque de développer l'événement principal jusqu'à la date de dernières nouvelles tant que cet événement n'a pas été observé.
- En présence de *K* événements compétitifs, si l'hypothèse d'indépendance n'est pas vérifiée, la méthode *Ignore* conduit à une surestimation de l'incidence cumulée spécifique.
- En présence d'événements compétitifs, la survie sans événement ne peut pas être calculée à partir du produit des survies spécifiques obtenues par la méthode *Ignore* et doit être estimée en considérant le délai jusqu'au premier des événements.

Méthode *Censure*

Principe

Avec la méthode *Censure*, seul le premier événement est pris en compte. En présence de risques compétitifs, si celui-ci ne correspond pas à l'événement principal d'intérêt, les données du sujet sont alors censurées à la date d'apparition de ce premier événement.

Définition des délais et des événements

Le temps de participation de chaque sujet correspond au délai de survie utilisé pour le calcul de l'estimation de la survie sans événement puisqu'il s'agit alors du délai de réalisation « du premier des *K* événements ». Concernant les sujets pour qui un événement est observé, il s'agit du délai entre la date d'origine et la date d'observation du premier événement quel qu'il soit. En revanche, si le premier événement n'est pas l'événement principal d'intérêt, les délais pour ces sujets sont censurés à la date d'apparition du premier événement. Lorsqu'aucun événement n'est observé, les délais sont censurés à la date de dernières nouvelles.

Estimation des survies et incidences cumulées spécifiques

Avec la méthode *Censure*, un événement autre que l'événement principal conduit à censurer les délais si cet événement n'est pas le premier. Ainsi, les événements compétitifs sont considérés

comme de vraies censures. Il est important de rappeler à ce stade que les techniques classiques d'analyse de données de survie supposent que la censure est non informative : la censure à un instant *t* n'apporte pas d'information sur l'incidence éventuelle d'événements ultérieurs. Cette indépendance entre le processus de censure et la survenue d'événements ultérieurs doit donc pouvoir être justifiée si cette méthode est retenue.

En censurant les événements à risque compétitif, la probabilité de survie spécifique à l'événement principal obtenue par cette approche correspond à la probabilité d'être indemne de l'événement principal *en tant que premier événement*. Cette probabilité sera différente de la probabilité de survie spécifique si l'hypothèse d'indépendance entre les événements ne peut être justifiée, et les estimations seront alors biaisées. En particulier, la survie spécifique marginale sera sous-estimée et l'incidence cumulée associée surestimée.

Dans le cas d'événements exclusifs, il est intéressant de noter que les méthodes *Ignore* et *Censure* sont indiquées conduisant aux mêmes estimations de l'incidence cumulée et de la survie spécifique.

Estimation de la survie sans événement

À l'inverse de la méthode *Ignore*, la fonction de survie sans événement peut être estimée par le produit des fonctions de survie spécifique $S_K^{cen}(t)$: $SSE(t) = S_1^{cen}(t) * S_2^{cen}(t) * \ldots * S_K^{cen}(t)$, même en présence d'événements dépendants.

Exemple fil rouge

Considérons les 5 patients de la *figure 1* ainsi que la paire de variables (t_i, c_i) où t_i correspond au temps jusqu'à la rechute locorégionale et c_i est une variable indicatrice égale à 1 si cette rechute est observée ou 0 si les données sont censurées. Avec la méthode *Censure*, les données sont donc censurées si les sujets sont perdus de vue avant l'observation d'une rechute locorégionale ou en cas d'apparition de métastases avant une rechute locorégionale. Ainsi :

- sujet 1 : aucun événement n'a été observé. Ses données sont censurées à la date de dernières nouvelles, ainsi $(t_1, c_1) = (12, 0)$;
- sujet 2 : la rechute locorégionale L est observée à 5 mois, ainsi $(t_2, c_2) = (5, 1)$;
- sujet 3 : des métastases M sont apparues au 3e mois. La méthode *Censure* implique que les délais de ce sujet sont censurés à l'apparition de ce premier événement qui n'est pas l'événement d'intérêt principal, ainsi $(t_i, c_i) = (3, 0)$;
- sujet 4 : la rechute locorégionale L est observée à 4 mois, ainsi $(t_i, c_i) = (4, 1)$;
- sujet 5 : des métastases M sont apparues au 2e mois avant la rechute locorégionale à 8 mois, ainsi $(t_5, c_5) = (2, 0)$.

Les fonctions de risque instantané h_k^{cen}, d'incidence cumulée I_k^{cen} et de survie spécifique S_k^{cen} peuvent être calculées à chaque instant *t* pour la rechute locale *(k = L) (tableau III)*.

Tableau III. Application de la méthode *Censure* (rechutes locorégionales).					
T (Sujet)	**Événement**	**À risque**	$h_L^{cen}(t)$	$1 - h_L^{cen}(t)$	$S_L^{cen}(t)$
0	–	–	–	1	1
4 (S4)	1	3	1/3	2/3	2/3 = 67 %
5 (S2)	1	2	1/2	1/2	1/3 = 33 %

Tableau III bis. Application de la méthode *Censure* (métastases).					
T (sujet)	**Événement**	**À risque**	$h_M^{cen}(t)$	$1 - h_M^{cen}(t)$	$S_M^{cen}(t)$
0	–	–	–	1	1
2 (S5)	1 (S5)	5	1/5	4/5	4/5 = 80 %
3 (S3)	1 (S3)	4	1/4	3/4	3/5 = 60 %

Ces probabilités peuvent être estimées pour l'apparition de métastases *(tableau III bis)*.

En appliquant la méthode *Censure*, les survies sans rechute locale et sans métastase à 12 mois sont estimées à 33 % et 60 % respectivement. Cette méthode permet d'estimer la *SSE* à partir du produit des survies de chacun des événements compétitifs, ainsi à 12 mois :

$$SSE(12) = S_L^{cen}(12) \times S_M^{cen}(12) = 1/3 \times 3/5 = 1/5 .$$

Avec la méthode *Censure*

• Tout premier événement autre que l'événement principal censure l'observation de celui-ci. Les données des sujets pour qui un événement autre que l'événement principal survient en premier sont censurées et confondues avec les « vraies » censures.
• En présence de *K* événements compétitifs, si l'hypothèse d'indépendance n'est pas vérifiée, la méthode *Censure* conduit à une surestimation de l'incidence cumulée spécifique.
• En présence d'événements compétitifs, la survie sans événement correspond au produit des survies spécifiques.

Méthode *Inclure*

Les méthodes *Ignore* et *Censure* produisent des estimations non biaisées de l'incidence cumulée, des survies spécifiques et sans événement uniquement sous l'hypothèse d'indépendance entre événements compétitifs. Cette hypothèse est rarement vraie et souvent difficilement vérifiable. Une méthode d'estimation adaptée à la présence de risques compétitifs est donc nécessaire.

Principe

Inclure les autres événements consiste à employer une méthodologie d'analyse des risques compétitifs où l'incidence spécifique de chaque type d'événement est estimée en présence des autres types d'événements. Les événements autres que l'événement principal ne sont ni ignorés ni censurés, mais considérés comme en compétition avec l'événement principal. Seule la « vraie » censure est définie comme un événement non encore observé.

Définition des délais et des événements

Le temps de participation de chaque sujet correspond au délai jusqu'à la date d'apparition du premier événement (tout type confondu) ou, à défaut, la date de dernières nouvelles. Avec les méthodes *Ignore* et *Censure*, une variable binaire permet d'indiquer si l'événement principal est observé $(c = 1)$ ou non, c'est-à-dire si les données sont censurées (c = 0). Dans le cas de l'approche *Inclure*, il s'agit d'une variable catégorielle indiquant le type d'évènement ou la censure. Ainsi, en présence de $K = 2$ événements compétitifs, cette variable indicatrice peut prendre trois valeurs : $e = 0$ (vraie censure), $e = 1$ (événement de type 1) ou $e = 2$ (événement de type 2).

Estimation des fonctions de survie et d'incidence cumulée spécifiques

L'estimation de la fonction d'incidence cumulée spécifique selon la méthode *Inclure* s'effectue en trois étapes. Au cours de la première étape, la fonction de survie sans événement *SSE(t)* est estimée à l'aide de la méthode de Kaplan-Meier. Dans la seconde étape, la fonction de risque spécifique à l'événement de type *k* est estimée. Enfin, la fonction d'incidence cumulée spécifique est estimée à partir de ces deux quantités.

Étape 1 : estimation de la survie sans événement *SSE(t)*

• Le temps de participation de chaque sujet est défini par le délai entre la date d'origine et la date d'apparition du premier événement tous types confondus (en cas d'événement) ou, à défaut, la date de dernières nouvelles (données censurées).

• À partir de ces définitions, la survie sans événement est estimée en utilisant la méthode de Kaplan-Meier.

Étape 2 : estimation du risque spécifique $h_k^{cen}(t)$ associé à l'événement *k*

• Pour ce calcul, le temps de participation et les événements pris en compte correspondent à ceux définis pour la méthode *Censure*. Le temps de participation correspond au délai d'apparition du premier événement quel qu'il soit et les données sont censurées uniquement s'il s'agit d'un événement autre que l'événement *k*.

• À partir de ces définitions, la fonction de risque spécifique $h_k^{cen}(t)$ est alors estimée par la méthode de Kaplan-Meier.

Étape 3 : estimation de l'incidence cumulée spécifique pour l'événement *k*

L'incidence cumulée spécifique pour l'événement *k* correspond à la probabilité que le premier événement ait lieu avant l'instant *t* et que ce premier événement soit du type *k*.

Soit *SSE(t-1)*, la survie sans événement juste avant l'instant *t*. Le risque d'observer l'événement *K* en premier événement au temps *t* correspond à la probabilité d'être indemne de tout événement au temps *(t-1)* et d'observer l'événement *k* au temps *t*. Statistiquement, ce risque instantané se calcule donc par le produit de la survie sans événement à l'instant *(t-1)* et du risque instantané de présenter l'événement *k* à l'instant *t* : $SSE(t-1)*h_k^{cen}(t)$. Cette probabilité instantanée peut donc être calculée à chaque instant *t*. L'incidence cumulée spécifique à l'événement *k* à l'instant *t* correspond alors à la somme des probabilités instantanées antérieures :

$$I_k^{inc}(t) = \sum_{\tau<t} SSE(\tau-1)h_k^{cen}(\tau)$$

Estimation des fonctions de survie sans événement et d'incidence cumulée (tout événement confondu)

Le calcul des incidences cumulées spécifique est effectué pour chacun des *K* événements compétitifs. L'incidence cumulée, tous événements confondus, correspond alors à la somme des incidences individuelles : $I^{inc}(t) = I_1^{inc}(t) + I_2^{inc}(t) + \ldots + I_K^{inc}(t)$.

Le complément de l'incidence cumulée correspond alors à la survie sans événement $SSE(t) = 1 - I^{inc}$.

Exemple fil rouge

Dans l'exemple de la *figure 1*, la rechute locorégionale et l'apparition de métastases sont les deux seuls événements compétitifs. Avec la méthode *Inclure*, un indicateur d'événement est associé à chaque type de premier événement. Ainsi, les deux variables d'analyse pour un sujet *i* sont donc (t_i, c_i), où $t_i \geq 0$ correspond au délai jusqu'au premier événement et $e_i = 1, \ldots, K$ est une variable indicatrice du type du premier événement. Dans l'exemple, cet indicateur peut prendre trois valeurs : 0 en cas d'absence d'événement (vraie censure), 1 si une rechute locale est observée en premier, et 2 s'il s'agit de métastase. Ainsi :

- sujet 1 : pas d'événement dans les 12 premiers mois de suivi, ainsi $(t_1, e_1) = (12, 0)$;
- sujet 2 : le 1er événement observé est L à 5 mois, ainsi $(t_2, e_2) = (5, 1)$;
- sujet 3 : le premier événement observé est M à 3 mois, ainsi $(t_3, e_3) = (3, 2)$;
- sujet 4 : le premier événement observé est L à 4 mois, ainsi $(t_4, e_4) = (4, 1)$;
- sujet 5 : le premier événement observé est M à 2 mois, ainsi $(t_5, e_5) = (2, 2)$.

Dans le *tableau IV*, nous présentons les estimations des risques instantanés h_L^{cen}, des incidences cumulées spécifiques I_L^{cen} ainsi que de la survie spécifique S_L^{cen}, relatifs à la rechute locorégionale,

Tableau IV. Application de la méthode *Inclure* (rechutes locales).						
Temps	**Rechutes locales**			***SSE***	I_L^{inc}	S_L^{inc}
	Événement	**À risque**	h_L^{cen}			
0	0	5	0	1	0	1
1	0	5	0	1	0	1
2	0	4	0	4/5	0	1
3	0	4	0	3/5	0	1
4	1	3	1/3	2/5	1/5	4/5
5	1	2	1/2	1/5	2/5	3/5
6	0	1	0	1/5	2/5	3/5
7	0	1	0	1/5	2/5	3/5
8	0	1	0	1/5	2/5	3/5
9	0	1	0	1/5	2/5	3/5
10	0	1	0	1/5	2/5	3/5
11	0	1	0	1/5	2/5	3/5
12	0	1	0	1/5	2/5	3/5

à partir de la méthode *Inclure*. Dans un premier temps, il est nécessaire d'estimer la survie sans événement en considérant le délai d'apparition du premier événement tous types confondus (en cas d'événement) ou la date de dernières nouvelles (données censurées) *(tableau I)*.

La seconde étape de la méthode *Inclure* consiste à estimer les risques instantanés spécifiques. Une rechute locorégionale n'est observée en tant que premier événement qu'à 4 mois. Les risques instantanés aux instants précédents sont donc nuls :

$$h_L^{cen}(0) = h_L^{cen}(1) = h_L^{cen}(2) = h_L^{cen}(3) = 0$$

À 4 mois, parmi les 3 sujets encore à risque (c'est-à-dire pour lesquels aucun événement n'a été observé jusque-là), une rechute locorégionale est observée, conduisant à un risque instantané à $t = 4$ mois de $h_L^{cen}(4) = 1/3$. De même, le risque instantané de rechute locorégionale est estimé à $h_L^{cen}(5) = 1/2$ à 5 mois et est nul au-delà.

Enfin, l'incidence cumulée spécifique est calculée à chaque instant t par la somme des produits des risques instantanés au temps t et de la survie sans événement au temps $(t - 1)$. Ainsi, pour la rechute locorégionale, l'incidence cumulée spécifique est donnée par :

$$I_L^{inc}(t) = \sum_{r<t} SSE(t-1) h_L^{cen}(t) = SSE(3) \times h_L^{cen}(4) + SSE(4) \times h_L^{cen}(5) = 2/5$$

En adoptant la même démarche pour l'apparition de métastases, l'incidence cumulée spécifique est obtenue de la façon suivante *(tableau IV bis)* :

$$I_M^{inc}(t) = \sum_{r<t} SSE(t-1) h_M^{cen}(t) = SSE(1) \times h_M^{cen}(2) + SSE(2) \times h_M^{cen}(3) = 2/5$$

L'incidence cumulée, tout événement confondu, correspond donc à la somme des incidences cumulées spécifiques, c'est-à-dire :

$$I^{inc}(t) = I_L^{inc}(t) + I_M^{inc}(t) = 4/5.$$

Tableau IV bis. Application de la méthode *Inclure* (métastases).

Temps	Métastases			*SSE*	I_M^{inc}	S_M^{inc}
	Événements	À risque	h_M^{cen}			
0	0	5	0	1	0	1
1	0	5	0	1	0	1
2	1	5	1/5	4/5	1/5	4/5
3	1	4	1/4	3/5	2/5	3/5
4	0	3	0	2/5	2/5	3/5
5	0	2	0	1/5	2/5	3/5
6	0	1	0	1/5	2/5	3/5
7	0	1	0	1/5	2/5	3/5
8	0	1	0	1/5	2/5	3/5
9	0	1	0	1/5	2/5	3/5
10	0	1	0	1/5	2/5	3/5
11	0	1	0	1/5	2/5	3/5
12	0	1	0	1/5	2/5	3/5

Méthode *Inclure*

- L'incidence spécifique de chaque type d'événement est estimée en présence des autres types d'événements.
- En présence de *K* événements compétitifs, seules les données des patients pour lesquels aucun événement n'a été observé sont censurées.
- En présence d'événements compétitifs, la survie sans événement à l'instant *t* est décomposée en estimant la proportion de chacun des événements. Des tests et modèles adaptés permettent de tester l'impact de facteurs pronostiques sur les survies spécifiques.

Comparaison des trois méthodes

En présence d'événements compétitifs, les incidences cumulées associées à chaque type d'événement peuvent être estimées selon trois approches différentes. Si ces événements sont dépendants et/ou non exclusifs, le temps de participation des sujets ainsi que la variable indicatrice de censure diffèrent selon la méthode retenue pour l'analyse statistique. Nous avons résumé dans le *tableau V* ces deux informations concernant les 5 sujets de l'exemple, selon que la méthode *Ignore*, *Censure* ou *Inclure* soit retenue.

Les méthodes *Ignore*, *Censure* et *Inclure* produiront des estimations identiques et non biaisées uniquement si les événements considérés sont indépendants. Dans les autres situations, les estimations seront différentes, voire biaisées pour les méthodes *Ignore* et *Censure*.

Avec la méthode *Ignore*, un sujet reste considéré à risque pour le critère principal tant que celui-ci n'a pas été observé, qu'un événement compétitif ait été observé antérieurement ou non. De plus, on suppose que le risque d'apparition de l'événement principal n'est pas affecté par l'incidence d'un événement compétitif. Lorsque les événements compétitifs sont indépendants, les estimations ne sont pas affectées par cette hypothèse, mais celles-ci seront biaisées dans le cas contraire.

En censurant les délais pour les sujets dont l'événement premier n'est pas celui d'intérêt, la méthode *Censure* répond à une question différente en s'intéressant à l'occurrence de l'événement principal en tant que *premier* événement. Il est cependant important de rappeler l'hypothèse fondamentale de censure non informative puisque l'on suppose alors que l'occurrence d'un événement ne modifie pas le risque d'observer un événement de type différent. Elle suppose également l'indépendance entre les différents événements. Ces hypothèses sont rarement vraies ou du moins difficilement vérifiables. Si elles ne sont pas satisfaites, les estimations obtenues par la méthode *Censure* seront biaisées.

La méthode *Inclure* permet de prendre en compte des événements à risque compétitif, qu'ils soient non exclusifs ou dépendants. Cette approche ne repose par sur l'hypothèse d'indépendance entre les différents événements. De plus, par l'utilisation d'un indicateur de type

Tableau V. Temps d'observation et indicateur d'événement pour les différentes méthodes.

	Temps de participation et indicateur d'événement		
	Ignore	***Censure***	***Inclure***
Patient 1	(12, 0)	(12, 0)	(12, 0)
Patient 2	(5, 1)	(5, 1)	(5, 1)
Patient 3	(9, 0)	(3, 0)	(3, 2)
Patient 4	(4, 1)	(4, 1)	(4, 1)
Patient 5	(8, 1)	(2, 0)	(2, 2)

d'événement, chaque type d'événement est pris en compte. Contrairement à la méthode *Censure*, les délais censurés représentent de vraies censures et permettent ainsi de garantir l'hypothèse de censure non informative. Enfin, avec la méthode *Inclure*, l'incidence cumulée tout événement confondu (et ainsi la survie sans événement) peut être décomposée en termes de probabilité de survenue de chacun des événements au cours du temps.

Prise en compte de variables explicatives

Il peut être intéressant de comparer des données de survies en fonction des caractéristiques des sujets. Lorsque la méthode *Censure* ou *Ignore* est retenue, le test du log-rank permet de comparer des courbes de survie (*cf.* chapitre III.1, page 129). De même, le modèle à risques proportionnels de Cox est utilisé dans le cadre de plusieurs variables explicatives (*cf.* chapitre IV.2 « Modèle de Cox et index pronostique », page 213).

Des tests et des modèles ont été spécifiquement développés dans le cadre d'une prise en compte appropriée des risques compétitifs par la méthode *Inclure*. Le test de Gray permet de comparer les fonctions d'incidence cumulée spécifiques en fonction d'une variable explicative [3]. Lorsque plusieurs variables sont à prendre en compte, le modèle de Fine et Gray est une adaptation du modèle de Cox permettant de réaliser des analyses multivariées en présence de risques compétitifs et ainsi de modéliser les fonctions d'incidences cumulées et de survie spécifique en fonction de plusieurs covariables [4]. Ce modèle permet de prendre en compte des événements compétitifs dépendants et/ou non exclusifs. Pour chaque type d'événement, un s*ub Hazard ratio* est estimé en fonction des différentes variables explicatives. Un s*ub Hazard ratio* supérieur à 1 indique alors qu'une variable explicative est associée à une augmentation relative du risque spécifique et ainsi à une incidence cumulée spécifique supérieure.

Enfin, dans le cadre d'un essai thérapeutique, l'incidence élevée d'événements compétitifs peut avoir un impact sur le nombre d'événements d'intérêt observés. Pour remédier à ce problème, différentes méthodes de planification ont par ailleurs été proposées afin de prendre en compte les risques compétitifs dans le calcul de la taille d'échantillon [5].

Quelques exemples

Événements exclusifs

Kutikov *et al.* [6] se sont intéressés à la survie de patients traités par chirurgie pour un cancer du rein localisé sans envahissement ganglionnaire. Cette population est généralement âgée avec de nombreuses comorbidités et peu de données sont disponibles concernant le bénéfice en termes de survie globale.

À partir de la base de données américaine du SEER *(Surveillance, Epidemiology, and End Result)*, les auteurs ont considéré trois événements : le décès lié au cancer du rein, le décès suite à un autre type de cancer et le décès d'une autre cause. Si les sujets n'étaient pas décédés au terme du suivi ou s'ils étaient perdus de vue en cours d'étude, leurs données étaient censurées à la date de dernières nouvelles.

Dans un premier temps, la survie globale a été analysée à partir de la méthode de Kaplan-Meier et le test du log-rank a été appliqué afin de mettre en évidence des associations éventuelles avec différents facteurs pronostiques, comme par exemple le sous-type histologique ou la taille de la tumeur. Les auteurs ont ainsi mis en évidence une association entre l'âge et la mortalité. De plus, toutes causes confondues, le risque de décès était plus important pour les hommes que les femmes. Puis, les auteurs se sont intéressés aux facteurs associés aux différentes causes de décès. Pour cela, un modèle de Fine et Gray a été adopté afin de prendre en compte les événements compétitifs et plusieurs variables explicatives. Pour chaque type d'événement, le *sub hazard ratio* a été estimé pour chaque variable. Les auteurs ont ainsi mis en évidence une association entre l'âge et la mortalité non liée au cancer du rein. Les tumeurs de taille importante étaient un facteur pronostique de décès du cancer du rein, mais inversement associé aux décès non dus au cancer. Ces résultats suggèrent qu'il est important de considérer les différentes causes de mortalité dans cette population particulière de sujets âgés. Lors du choix du traitement, la prise de décision doit prendre en compte toutes les causes de décès éventuelles. Cependant, l'application à ces types de données implique nécessairement que les causes de décès soient connues pour tous les sujets.

Événements non exclusifs

L'objectif principal du traitement par radiothérapie est le contrôle local. À la suite du traitement, le patient est à risque de différents types d'événements, notamment les récidives locales, les métastases à distance et les complications à long terme. Dans le cas où l'on s'intéresse aux complications à long terme, la méthode *Ignore* n'est pas adaptée. En effet, il faudrait déterminer si les complications à long terme après une rechute sont liées au traitement initial de la maladie ou au traitement de la rechute. Il existe un débat récurrent dans la littérature [7] sur le choix d'une des deux autres méthodes : *Censure* ou *Inclure*. Par exemple, un patient qui a reçu une radiothérapie intensive qui a éradiqué la tumeur peut souffrir de complications à long terme. Certains auteurs soutiennent que l'utilisation de la méthode *Inclure* va dénaturer le risque car de nombreux patients atteints d'une maladie agressive peuvent décéder avant l'apparition d'une complication tardive. Ils préconisent d'utiliser la méthode *Censure* pour estimer l'incidence cumulée des complications tardives. Mais cette méthode de calcul va surestimer le taux réel de complication tardive. Si le traitement avait été appliqué à des patients à faible risque, l'utilisation de la méthode *Censure* pourra mettre en évidence une incidence supérieure des taux de complication par rapport à la méthode *Inclure*.

Conclusions

En présence d'événements compétitifs, il est nécessaire, avant toute analyse, d'étudier les relations éventuelles entre ces événements : sont-ils indépendants ? À défaut, sont-ils exclusifs ? En effet, plusieurs méthodes permettent d'analyser les risques compétitifs, et les estimations peuvent différer selon les associations qui existent entre ces événements.

En effet, en présence d'événements indépendants, que l'on ignore ou que l'on censure les données en cas de survenue d'un événement compétitif n'affectera pas les estimations d'incidence et de survie spécifiques. En cas de dépendance, les estimations de l'incidence cumulée spécifique seront surestimées. Dans ce contexte particulier, il convient d'adopter une méthodologie adaptée.

Références

1. Bentzen SM, Dorr W, Anscher MS, *et al*. Normal tissue effects: Reporting and analysis. *Semin Radiat Oncol* 2003 ; 13 (3) : 189-202.

2. Com-Nougué C, Guérin S, Rey A. Assessment of risks associated with multiple events. *Rev Epidemiol Sante Publique* 1999 ; 47 (1) : 75-85.

3. Fine J, Gray R. A proportional hazards model for the subddistribution of a competing risk. *J Am Stat Assoc* 1999 ; 94 : 496-509.

4. Gray R. A class of K sample tests for comparing the cumulative incidence of a competing risk. *Ann Stat* 1988 ; 16 : 1141-54.

5. Kutikov A, Egleston BL, Wong Y, Uzzo R. Evaluating overall survival and competing risks of death in patients with localized renal cell carcinoma using a comprehensive nomogram. *J Clin Oncol* 2010 ; 28 (2) : 311-7.

6. Latouche A, Porcher R. Sample size calculations in the presence of competing risks. *Stat Med* 2007 ; 26 (30) : 5370-80.

7. Tai BC, Machin D, White I, Gebski V. Competing risks analysis of patients with osteosarcoma: A comparison of four different approaches. *Stat Med* 2001 ; 20 (5) : 661-84.

Suivi et surveillance

T. Filleron, M. Gauthier

Lors de l'analyse des données de survie, la première question à se poser est de connaître la durée du suivi, car les estimations de survie à 3 ou à 5 ans n'ont pas la même précision si les patients ont un suivi très court ou très long. En effet, pour les données de survie, la précision dépend du nombre de patients, du nombre d'événements mais également de la durée du suivi. Il est donc nécessaire, avant toute analyse de survie, de quantifier le suivi. Les différentes méthodes pour estimer ce paramètre vont être abordées dans ce chapitre.

Introduction

Le suivi des patients traités pour un cancer constitue un des facteurs de progrès essentiel de la cancérologie et de la médecine. L'objectif premier du suivi des patients est une perspective médicale, orientée vers la surveillance de la maladie traitée. Il s'agit en effet de déceler des récidives locorégionales, métastatiques ou un second cancer à un stade précoce afin que le traitement ait le plus de chances possibles d'obtenir un résultat bénéfique pour le patient. Pour cela, des recommandations de surveillance explicitant fréquence et modalités des examens (clinique, imagerie, marqueurs, etc.) sont proposés pour les différentes localisations cancéreuses par des sociétés savantes reconnues comme l'*American Society of Clinical Oncology* (ASCO) et l'*European Society of Medical Oncology* (ESMO) ou par des organismes nationaux (Haute Autorité de santé, Institut national du cancer en France). Cependant, cet objectif est loin de résumer les divers aspects de la période suivant le premier traitement ou la première stratégie thérapeutique du cancer. Le contenu du suivi d'un patient traité s'inscrit dans une démarche non exclusivement médicale et répond à d'autres impératifs ; il peut s'agir d'un suivi :

- *social :* prenant en compte la réalité de la réinsertion et l'éventuelle réhabilitation de l'individu ;
- *médico-scientifique :* tendant à évaluer le bien-fondé de l'affectation des moyens nécessaires à la prise en charge médicale d'un individu au sein de la masse globale de ressources disponibles.

Concernant ce dernier point, l'essai thérapeutique est l'étape indispensable qui permet de mesurer et/ou de comparer les conséquences des traitements tant sur le plan de l'efficacité que de la tolérance, avec notamment les séquelles et les complications. Dans les essais de phase II et III, les critères de jugement traditionnellement utilisés pour démontrer un bénéfice clinique pour une

nouvelle thérapeutique sont le taux de réponse et la survie globale. Des critères autres comme le temps jusqu'à progression et la survie sans progression sont de plus en plus utilisés car ils offrent différents avantages, notamment par rapport à la survie globale [1, 2] :

• la progression se situe avant le décès, ce qui permet de réduire le nombre de sujets à inclure et la durée de l'essai ;

• les traitements de rattrapage n'influencent pas la date de progression.

Cependant, ces critères de jugement peuvent être soumis à différents biais liés au suivi. Le biais d'évaluation (biais de mesure) est le plus connu : il survient lorsque le critère de jugement n'est pas évalué de la même manière dans les deux bras, par exemple si la fréquence des visites de suivi est différente. Dans ce chapitre, les différentes sources de biais liées au suivi dans les essais de phase II et III ou les études rétrospectives étudiant des durées de survie vont être abordées et commentées.

Dans la seconde partie, une autre question ayant trait au suivi est souvent posée au cours des présentations d'un essai clinique : « les données sont elles “matures” ? ». Pour répondre à cette question, plusieurs mesures sont utilisées dans la littérature : recul médian, suivi médian... Nous aborderons la signification de ces différents termes.

Rythme de suivi et essai

Le biais d'évaluation, appelé de façon plus générale biais de mesure, survient lorsque le critère de jugement n'est pas recherché de manière identique entre les groupes de traitements. Il peut être lié à des modalités d'examen différentes et/ou à une fréquence d'examens différente entre les groupes. L'objectif de cette section n'est pas de proposer un rythme et des méthodes d'évaluation optimale, mais de commenter différentes situations inhérentes au suivi (fréquences et/ou modalités) dans les essais thérapeutiques à l'aide d'un exemple fil rouge.

Modalités d'examens, fréquence des évaluations

Dans les essais comportant plusieurs bras, les modalités de suivi (fréquence et type d'examen) doivent être identiques entre les bras afin d'éviter les biais. En effet, dans le cadre d'un essai où l'objectif principal est défini par la survie sans rechute (récidive locorégionale, métastase), un rythme de suivi et/ou des examens différents entre les bras peuvent amener à détecter les événements de manière plus précoce dans un bras. C'est souvent le bras expérimental qui demande plus de suivi quand il est comparé à un bras sans traitement. Ce bras sera donc « désavantagé », ce qui peut avoir pour conséquence de mettre en évidence une différence statistiquement significative qui est uniquement liée au suivi.

Fréquence des évaluations et critères RECIST

Afin d'illustrer les différentes situations liées à la fréquence des évaluations au cours du suivi d'essai thérapeutique, un exemple « fil rouge » de ce chapitre va être présenté, puis différentes situations seront commentées.

Exemple « fil rouge »

Prenons l'exemple d'un essai de phase II, dont les critères principal et secondaire sont respectivement la survie sans progression et la durée de la réponse. Le temps jusqu'à progression est défini par le délai entre la date d'inclusion et la date de la progression ou la date de dernières nouvelles (données censurées) [3]. La durée de la réponse est définie, pour les patients qui répondent, comme étant le délai entre la date de la réponse et la date où une progression est documentée.

La *figure 1* représente le pourcentage de changement de la taille tumorale par rapport à l'évaluation tumorale initiale au cours du temps pour un patient. Si les évaluations étaient effectuées tous les jours, la date de la réponse partielle serait à 2 mois (rond). Les dates exactes du nadir et de la progression seraient respectivement à 5 mois (carré) et 12 mois (triangle).

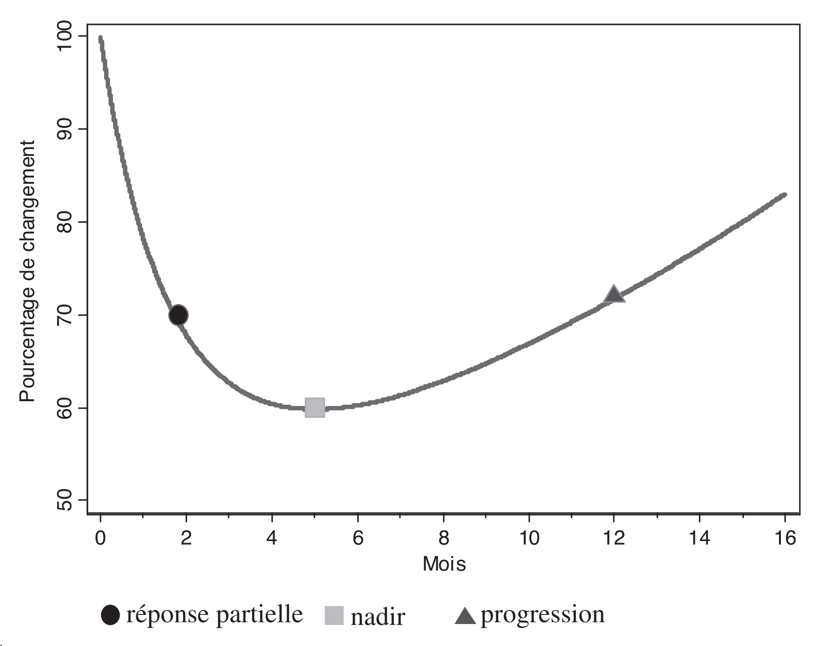


Figure 1. Évolution du changement par rapport à l'évaluation tumorale initiale.

Dans la « vraie vie »

Contrairement à l'exemple fil rouge, dans la « vraie vie » les dates exactes de la réponse, de la progression et du nadir ne sont jamais réellement connues, mais elles se situent entre deux visites de réévaluation. Généralement, la date imputée correspond à la date de la visite. Ces trois paramètres sont donc largement influencés par le rythme de suivi. De plus, la date du nadir a un impact sur la date de la progression car un patient est considéré en progression s'il y a une augmentation de 20 % par rapport au nadir selon les critères RECIST [3].

Cinq calendriers de suivi vont être appliqués à l'exemple fil rouge pour illustrer l'impact de la fréquence de réévaluation sur la survie sans progression et la durée de la réponse : C1 : tous les mois ; C2 : tous les 2 mois ; C3 : tous les 3 mois ; C4 : tous les 4 mois ; et C5 : alterné : examen à 2 mois puis tous les 3 mois.

Le calendrier 1 correspond au calendrier idéal : en effet, le véritable nadir correspond à une date de visite *(tableau I)*. La réponse partielle est observée à 2 mois, le nadir à 5 mois et la progression à 12 mois. Le délai jusqu'à progression et la durée de la réponse sont respectivement de 12 et 10 mois. Pour le calendrier 2, la première réponse partielle se situe à 2 mois mais le nadir est à 6 mois, la progression est alors documentée à 14 mois, ce qui correspond à une durée de la réponse de 12 mois.

Si un raisonnement similaire est effectué pour l'ensemble des calendriers, le « véritable » temps jusqu'à progression est de 12 mois alors que, selon le calendrier de suivi, il peut varier entre 12 mois et 16 mois. Similairement, la durée de la réponse varie entre 10 et 12 mois.

En résumé

La définition du rythme et des modalités de suivi (examens cliniques, imagerie, marqueurs tumoraux) est primordiale dans la conception d'un essai clinique. Des délais importants entre les évaluations peuvent conduire à passer à côté d'une différence significative entre les traitements, mais des évaluations plus fréquentes augmentent la charge sans nécessairement améliorer l'exactitude de la date de progression. Si les traitements en cours d'évaluation sont administrés à des intervalles différents, il est nécessaire de prévoir, dans le protocole, des dates d'évaluation identiques entre les bras. En effet, dans le cas contraire, le bénéfice observé pourrait être lié à la fréquence des évaluations. De plus, dans le choix de la fréquence des évaluations, il est important de prendre en considération les rythmes utilisés dans d'autres études sur la localisation cancéreuse concernée. La fréquence des évaluations doit aussi être compatible avec la pratique clinique pour les suivis des patients traités pour la maladie.

Le rythme de suivi et les modalités de surveillance doivent être planifiés dans le protocole de l'essai de façon à ce que chaque étape soit définie à l'avance pour éviter au maximum les biais. En effet, la valeur scientifique d'un essai thérapeutique dépend autant de la qualité du suivi des malades (et des données recueillies) que de la stratégie de départ. Une analyse statistique mal

Tableau I. Rythme de suivi et évaluation.						
		C1	**C2**	**C3**	**C4**	**C5**
1		S				RP
2	RP	RP	RP			
3		RP		RP		
4		RP	RP		RP	
5	**Nadir**	RP				RP
6		RP	RP	RP		
7		RP				
8		RP	RP		RP	RP
9		RP		RP		
10		RP	RP			
11		RP				RP
12	MP	MP	RP	RP	RP	
13		MP				
14		MP	MP			MP
15		MP		MP		
16		MP	MP		MP	
DP	12	12	14	15	16	14
DR	10	10	12	12	12	12

DP : délai jusqu'à progression ; DR : durée de la réponse ; MP : maladie progressive ; RP : réponse partielle. Le nadir est en gras.

conduite pourra toujours être refaite, des données manquantes peuvent parfois être récupérées, mais un protocole mal suivi ou une stratégie de suivi inadéquate sont des erreurs fatales quant à la valeur des conclusions.

Rythme et suivi

Le rythme et les modalités de suivi doivent :
- planifiés dans le protocole ;
- être identiques dans les différents bras pour les essais randomisés ;
- être compatibles avec la pratique clinique ;
- prendre en compte les rythmes des autres études.

Estimation du suivi

Dans la littérature, plusieurs notions de suivi sont utilisées car il n'existe pas de consensus pour ce point primordial à la fiabilité des estimations des courbes de survie. En effet, plusieurs revues de la littérature ont présenté les différentes méthodes utilisées ainsi que leurs limites [4-6].

Définitions du suivi *(follow-up)*

Lorsque le *follow-up* est évoqué, il est nécessaire de parler du suivi, qui correspond au temps pendant lequel le patient a été suivi, ou du recul, qui correspond au temps passé depuis que le patient a été inclus dans l'étude *(figure 2)*.

Ces notions de suivi et recul sont souvent décrites à l'aide de la médiane. Cependant, il existe de nombreuses méthodes pour les définir. Les principales méthodes retrouvées dans les différentes revues de la littérature sont les suivantes :
- le suivi est calculé seulement sur les délais de suivi des patients encore en vie ;
- le temps minimum de suivi est déterminé ;
- le suivi correspond au délai entre la date d'origine (par ex. date de diagnostic) et le décès ou la date de dernières nouvelles ;
- le suivi correspond au délai entre la date d'origine (par ex. date de diagnostic) et la date de fin d'étude.

Méthodes de calcul et leurs limites

Plusieurs méthodes de calcul sont utilisées ou suggérées par Schemper [7]. Pour cela, supposons une étude avec le recrutement des patients entre T_1 et T_2 et une analyse des données prévue à la date de fin d'étude notée T_3. Pour chaque individu i ($1 \leq i \leq n$), le temps d'entrée dans l'étude (*i.e* date d'inclusion) est notée t_{1i} et la date du dernier enregistrement t_{2i}. Le statut du patient S_i vaut 1 s'il est décédé au dernier enregistrement du patient ; il vaut 0 s'il est toujours vivant ou perdu de vue à la date de dernières nouvelles.

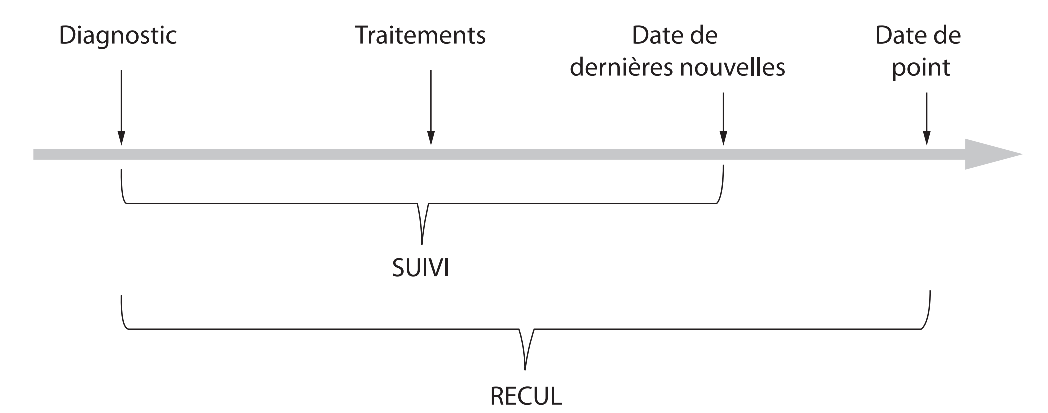


Figure 2. Suivi et recul depuis le diagnostic.

Bien qu'il soit plus pratique de décrire le suivi à l'aide d'une distribution empirique (médiane ou autres percentiles), plusieurs méthodes ont été décrites afin de permettre une quantification du suivi plus précise.

Temps d'observation des patients (TO)

Méthode de calcul

Le temps d'observation de tous les patients prend en compte les délais d'observation de tous les patients qu'ils soient en vie ou décédés ; il correspond au délai entre la date d'inclusion dans l'étude et la date de dernières nouvelles : $TO_i = t_{2\mathrm{i}} - t_{1\mathrm{i}}$.

Limites de la méthode

Le suivi calculé par cette méthode peut être sous-estimé dans le cas où le risque de décès est élevé au cours du temps. En effet, si deux études avec un suivi identique des patients mais avec une survie moins longue pour l'une d'elles sont comparées, le temps d'observation des patients sera inférieur pour la population d'étude qui a la survie la moins longue.

Temps d'observation chez les patients en vie (TC)

Méthode de calcul

Le temps d'observation chez les patients en vie se limite aux patients en vie à la date de dernières nouvelles (e temps d'observation des patients décédés est ignoré). Il est défini par le délai entre la date d'entrée dans l'étude et la date de dernières nouvelles :

$TC_i = t_{2\mathrm{i}} - t_{1\mathrm{i}}$ (seulement si $S_i = 0$)

Limites de la méthode

Le suivi calculé à l'aide de cette méthode est systématiquement sous-estimé car seuls les sujets censurés sont pris en compte dans le calcul. En effet, les patients perdus de vue ont la même probabilité de décéder que ceux qui termineront l'étude. Plus un patient est suivi longtemps, plus la probabilité qu'il décède au cours du suivi augmente. À l'inverse, plus la durée du suivi est courte, plus la probabilité que le patient soit en vie à la date de dernières nouvelles est élevée. Plus les patients seront suivis longtemps, moins le temps d'observation chez les patients en vie sera calculable car la probabilité qu'ils décèdent va augmenter. Avec un suivi long, moins il y aura de patients pris en compte dans le calcul du suivi médian.

Temps de recul (TR)

Méthode de calcul

Le recul est défini de la même manière pour tous les patients, qu'ils soient en vie ou décédés. Il s'agit du délai entre la date de fin d'étude et la date d'inclusion du patient : $TR_i = T_3 - t_{1\mathrm{i}}$.

Limites de la méthode

Cette fois-ci, la méthode de calcul a tendance à surestimer le temps de suivi car elle ne pénalise pas les données censurées. En effet, le temps de recul calculé correspond au temps écoulé entre l'entrée dans l'étude de chaque patient et une date de point qui est la même pour tous les patients qu'ils aient eu un suivi complet ou qu'ils soient perdus de vue en cours de suivi. Un patient perdu de vue rapidement après son inclusion aura le même poids dans le calcul qu'un patient qui a réalisé le suivi complet.

Fonction du temps connu (FTC)

Méthode de calcul

La fonction de temps connu est un composé des méthodes *TC* et *TR*. En effet, pour les patients vivant à la date de dernières nouvelles ($S_i = 0$), elle prend pour valeur TC. Alors que pour les patients décédés ($S_i = 1$), elle correspond à TR.

$FTC_i = t_{2i} - t_{1i}$ (si $S_i = 0$) et $FTC_i = T_{3i} - t_{1i}$ (si $S_i = 1$)

Limites de la méthode

Avec cette méthode, il est considéré que le statut pour un sujet décédé ne peut pas changer et il peut ainsi être affirmé que le patient a été suivi jusqu'à la fin de l'étude. Cette méthode a également tendance à surestimer le temps de suivi surtout lorsque le nombre de perdus de vue augmente.

Suivi selon la distribution de Korn

Méthode de calcul

Les quantiles pour la distribution de Korn sont obtenus à partir d'une fonction qui estime la probabilité *P* d'être en dessous du suivi au temps t' ($t'>0$) :

$P(t') = P(L>t' | E>t')\ P(E>t')$

où $P(E>t')$ est la proportion de sujets avec $TR_i > t'$, *L* correspond au temps jusqu'aux dernières nouvelles, $P(L>t' | E>t')$ est calculée à l'aide de la méthode de Kaplan-Meier inversé (voir paragraphe suivant) pour les sujets avec $TR_i \geq t'$.

Limites de la méthode

Contrairement aux quatre méthodes précédentes, celle-ci présente moins de limites car elle tient compte des différents risques des perdus de vue en fonction de l'entrée tardive ou non des patients dans l'étude. Elle permet notamment de décrire la distribution de la censure comme la distribution du suivi potentiel. Cependant, cette méthode est bien plus complexe à mettre en œuvre pour déterminer la valeur du suivi, car elle n'est pas développée dans les logiciels d'analyse classique. De plus, cette méthode, tout comme les méthodes TR et FTC, exige la connaissance d'une

date de fin d'étude T3 qui est souvent choisie de façon arbitraire compte tenu du fait que le statut du patient est connu individuellement à la date de sa dernière visite et non à une date de point commune à tous les patients.

Méthode de Kaplan-Meier inversé

Méthode de calcul

La méthode de calcul du temps de suivi est la même que celle utilisée pour l'estimation de la fonction de survie par la méthode Kaplan-Meier sauf que l'indicateur du statut est inversé. En effet, le décès ($S_i = 1$) sera alors considéré comme une censure, et la censure (perdu de vue ou patient vivant) deviendra l'événement. Le suivi médian correspond au point de la courbe situé à 50 %. Cette méthode a l'avantage de prendre en compte le suivi de tous les patients et de censurer les patients décédés au moment de leur décès.

Limites de la méthode

Cette méthode, qui prend en compte le concept du suivi de Korn, peut être calculée plus facilement avec des programmes informatiques facilement disponibles. Celle-ci reflète plus précisément les changements dans la qualité du suivi car elle prend en compte les temps de suivi de chaque patient. Cependant, la méthode de Kaplan-Meier inversé *(Reverse Kaplan-Meier)* suppose que la qualité de suivi ne dépend pas du rythme des inclusions des patients car elle calcule le temps écoulé entre la date d'origine et la date de dernières nouvelles pour chaque patient.

Ce mode de calcul n'est pas adapté si, par exemple, 8 patients sur 20 sont censurés assez tôt (par exemple à 3, 4, 5, 6, 8, 9, 10 et 11 mois) et 1 patient est censuré très tard (à 20 mois) alors que les décès sont observés à 1, 2, 7, de 12 à 19 mois pour les 11 autres patients. Dans ce cas, le suivi médian calculé par cette méthode sera de 20 mois alors que la médiane parmi les patients censurés est de 8 mois.

Exemple

Les différentes méthodes sont illustrées dans le cadre d'un exemple incluant 20 patients. La date d'ouverture de l'essai est le 10/01/2000. La date de fin d'étude est le 31/12/2009. Les données associées à chaque patient sont résumées dans le *tableau II* : la date d'inclusion, la date de l'événement ou la date de dernières nouvelles et l'indicateur d'événement. Par exemple, la date d'inclusion du patient 3 est le 15/06/2000, et il est perdu de vue le 14/07/2006. Le patient 20 est inclus dans l'étude le 03/04/2008 et est décédé le 25/07/2008.

Les délais TO, TC, TR et FTC présentés précédemment sont résumés dans le *tableau II*. La médiane de ces différentes quantités est donnée dans la dernière ligne du tableau (TO = 719,5 ; TC = 1 181 ; TR = 2 291,5 ; et FTC = 1 768,5). En utilisant la méthode de Kaplan-Meier inversé, la médiane de suivi est estimée à 1 606 jours.

Tableau II. Exemple de calculs des différentes quantités de suivi (en jours).							
Patient	**Date inclusion**	**Date du décès ou de dernières nouvelles**	**Événement**	**TO**	**TC**	**TR**	**FTC**
1	01/02/2000	15/02/2001	1	380		3 621	3 621
2	01/04/2000	31/12/2009	0	3 561	3 561	3 561	3 561
3	15/06/2000	14/07/2006	0	2 220	2 220	3 486	2 220
4	03/02/2001	01/01/2009	1	2 891		3 253	3 253
5	14/03/2001	05/02/2002	1	328		3 214	3 214
6	26/03/2001	01/04/2006	1	1 832		3 202	3 202
7	01/06/2001	03/02/2009	0	2 804	2 804	3 135	2 804
8	06/04/2002	02/01/2003	0	271	271	2 826	271
9	14/02/2003	01/02/2004	0	352	352	2 512	352
10	31/03/2003	03/05/2004	1	399		2 467	2 467
11	16/03/2004	23/06/2005	1	464		2 116	2 116
12	06/05/2004	06/08/2008	0	1 553	1 553	2 065	1 553
13	01/11/2004	26/03/2009	0	1 606	1 606	1 886	1 606
14	03/12/2004	05/12/2006	1	732		1 854	1 854
15	14/03/2005	01/06/2007	0	809	809	1 753	
16	23/05/2005	03/06/2009	1	1 472		1 683	1 683
17	12/01/2006	01/02/2006	0	20	20	1 449	20
18	23/06/2007	06/04/2008	1	288		922	922
19	24/01/2008	31/12/2009	0	707	707	707	707
20	03/04/2008	25/07/2008	1	113		637	637
			Médiane	**719,5**	**1 181**	**2 291,5**	**1 768,5**

TO : temps d'observation des patients, TC : observation chez les survivants ; TR : temps de recul ; FTC : fonction du temps connu.

Rôle du suivi médian

La précision des estimations des taux de survie à un temps donné dépend de la durée du suivi. Afin d'éviter une mauvaise interprétation des résultats, il est nécessaire de déterminer si le suivi est assez long par rapport à la maladie étudiée. Tout d'abord, les dates de point utilisées pour les analyses des données et le calendrier des évaluations ont un impact sur la probabilité d'observer

des événements liés aux délais. Ensuite, cette variabilité complémentaire peut influencer la validité des comparaisons entre plusieurs groupes, surtout quand une grande majorité de patients est vivante et sans événement. Enfin, le suivi médian est très utile dans l'interprétation des courbes de survie de Kaplan Meier, car il permet de définir les intervalles de temps où les estimations des taux de survie peuvent être considérées comme fiables. Il s'agit aussi d'un indicateur permettant d'évaluer la durée du suivi. Les points indispensables :

- définition de trois dates :
– date d'origine (inclusion, randomisation, diagnostic, etc.),
– date de fin pour le patient (date de survenue de l'événement, décès, date de dernière nouvelle si le patient est perdu de vue),
– date de point pour l'analyse ;
- description de la méthode de calcul ;
- dans un article : bien présenter ces deux points.

Suivi médian

- Le suivi médian et le recul médian ne sont pas synonymes.
- Le suivi médian est un indicateur de la durée du suivi, qui permet notamment de déterminer si la durée du suivi est adaptée à la localisation étudiée, il permet aussi de déterminer la stabilité des estimations des taux de survie au cours du temps.

Quantification de l'exhaustivité du suivi

Pour chaque individu, le temps de suivi correspond au temps écoulé entre la date d'origine et la date de survenue de l'événement s'il survient, sinon la date de dernières nouvelles si l'événement n'a pas été observé. Tous les patients sont inclus dans l'analyse de survie jusqu'à la dernière date connue : événement ou censure.

Dans un essai clinique, l'exhaustivité du suivi *(completeness of follow-up)* est particulièrement importante car des suivis différents entre les groupes peuvent avoir un impact sur l'analyse des résultats. Il est donc important de rapporter le nombre de patients perdus de vue ou ayant des reculs insuffisants dans une étude, car c'est un indicateur de données incomplètes. Il permet d'évaluer si les patients ont été suivis de la même manière dans les deux groupes. Cependant, il ne suffit pas de représenter correctement le « temps perdu ».

L'index d'exhaustivité du suivi (C) a été proposé par Clark, qui quantifie l'effet des perdus de vue [8]. Il s'agit du ratio entre la somme des « temps de suivi » observés (TO) et le temps total de suivi potentiel de l'étude (TR). Ce temps de suivi potentiel est défini par le temps entre l'entrée dans l'étude et la date de fin ou la date de l'événement s'il y a eu événement.

Comment calculer C ?

L'équation de la mesure de l'exhaustivité du suivi est la suivante :

$$C = 100 \cdot \sum_{i=1}^{n} t_i \Big/ \sum_{i=1}^{n} t_i^*$$

où t_i = temps de survie observé (TO) pour le nième patient, t_i^* = temps de suivi potentiel (*i.e.* le temps de recul TR) pour le nième patient, n = nombre de patients.

Le numérateur correspond à la somme des temps de suivi, sans tenir compte de l'événement.

Le dénominateur est la somme des temps de suivi potentiels.

Dans le cas de l'exemple du *tableau II*, le numérateur et le dénominateur sont respectivement égaux à 22 802 et 46 349, l'index de Clark *C* est donc de 49,2 %.

Quand calculer C ?

C peut être calculé dans les essais cliniques et dans les études de cohorte prospectives.

Comment interpréter C ?

Meilleure est la valeur de *C*, plus les données sont complètes, et plus fiables sont les résultats.

Comment et pourquoi utiliser C ?

• Dans les **essais cliniques**, *C* devrait être présenté pour chaque groupe de traitement afin de rechercher des différences entre les suivis de chaque traitement.

• Dans les études de **cohortes prospectives**, *C* devrait être présenté séparément pour le groupe de patients exposés ou pour le groupe non exposé et ce, dans le cas où le groupe des patients exposés est connu au début de l'étude.

• Dans les **études multicentriques**, *C* peut être utilisé pour vérifier l'homogénéité des suivis entre les centres.

Conclusion

L'exhaustivité du suivi est un indicateur important de la qualité de l'étude, mais il a tendance à être négligé dans les études de survie et peu clair dans les publications des essais cliniques.

Références

1. Buyse M, Burzykowski T, Carroll K, *et al*. Progression-free survival is a surrogate for survival in advanced colorectal cancer. *J Clin Oncol* 2007 ; 25 : 5218-24.

2. Michiels S, Le Maître A, Buyse M, *et al*. Surrogate endpoints for overall survival in locally advanced head and neck cancer: meta-analyses of individual patient data. *Lancet Oncology* 2009 ; 10 (4) : 341-50.

3. Eisenhauer EA, Therasse P, Bogaerts J, *et al*. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur J Cancer* 2009 ; 45 (2) : 228-47.

4. Mathoulin-Pelissier S, Gourgou-Bourgade S, Bonnetain F, Kramar A. Survival end point reporting in randomized cancer clinical trials: A review of major journals. *J Clin Oncol* 2008 ; 26 (22) : 3721-6.

5. Altman DG, De Stavola BL, Love SB, *et al*. Review of survival analyses published in cancer journals. *Br J Cancer* 1995 ; 72 : 511-8.

6. Shuster JJ. Median follow-up in clinical trials. *J Clin Oncol* 1991 ; 9 : 191-2.

7. Schemper M, Smith TL. A note on quantifying follow-up in studies of failure time. *Control Clin Trials* 1996 ; 17 : 343-6.

8. Clark TG, Altman DG, De Stavola BL. Quantification of the completeness of follow-up. *Lancet* 2002 ; 359 : 1309-10.

Partie IV
Analyses multivariées

Régression logistique et courbes ROC

C. Bascoul-Mollevi, A. Kramar

La régression logistique est une méthode très utilisée en épidémiologie et recherche clinique. Par rapport à la régression linéaire, qui caractérise une relation entre une variable dépendante quantitative et une ou plusieurs variables explicatives, la régression logistique caractérise une relation entre une variable dépendante qualitative comme la réponse, en général dichotomique, et une ou plusieurs variables explicatives qualitatives ou quantitatives. Par exemple, on peut expliquer la survenue d'un événement (malade/non malade) en fonction de facteurs cliniques, biologiques ou génétiques. La section intitulée « Le modèle logistique polytomique » présente une généralisation dans le cas d'une réponse en trois catégories. Les courbes ROC *(Receiver Operating Characteristic)*, souvent associées à la régression logistique, sont utiles dans le contexte de l'évaluation des tests diagnostiques et de dépistage, en particulier pour l'évaluation de nouveaux marqueurs dont la mesure est continue. Cette méthodologie a été initialement développée au cours de la Seconde Guerre mondiale pour la détection des signaux RADAR afin de distinguer le bruit de fond du signal et ainsi mieux détecter les avions ennemis.

Introduction

Les méthodes décrites dans ce chapitre seront illustrées par les données d'une étude évaluant l'utilisation d'un panel d'auto-anticorps dirigés contre des antigènes associés aux tumeurs en tant que technique de détection précoce des cancers du sein et, plus particulièrement, des carcinomes *in situ* (CIS) [1] ; en effet, le dépistage par mammographie a permis d'augmenter la détection des cancers du sein à des stades précoces. La réactivité sérique du panel constitué par ces 5 auto-antigènes – la cyclophilline A (PPIA), la peroxirédoxine 2 (PRDX2), l'immunophiline FKBP52, la *heat shock protein* 60 (HSP60) et la mucine 1 (MUC1) – a été testée sur 235 échantillons de sérums (93 contrôles, 82 CIS et 60 carcinomes infiltrants) en utilisant des tests ELISA spécifiques *(cf. figure 1)*.

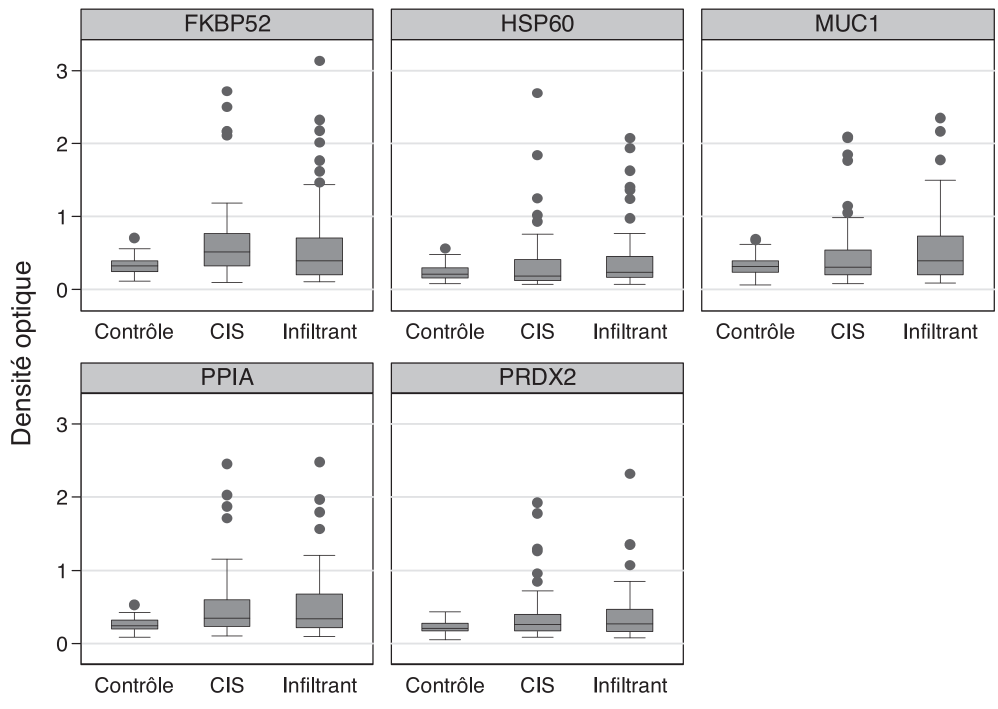


Figure 1. Distribution des auto-antigènes. La boîte est délimitée par les 25e et 75e percentiles et contient 50 % des observations ; médiane (trait horizontal à l'intérieur de la boîte) ; 5e et 95e percentiles (traits fins horizontaux) ; symboles (valeurs extrêmes).
CIS : carcinome *in situ*.

Régression logistique

Cette section présente la régression logistique simple (avec une seule variable explicative), la régression logistique multiple (avec plusieurs variables explicatives) et la régression polytomique (avec la variable réponse en plusieurs catégories).

Le modèle logistique simple

La régression logistique est fondée sur la modélisation de la probabilité de la survenue ou non de l'événement Y, en fonction de variables explicatives notées X_i $(i = 1, \ldots, p)$ par une fonction logistique. En d'autres termes, dans le cas d'une seule variable explicative $(p = 1)$, le modèle s'écrit :

$$P(Y=1 \mid X) = \frac{\exp(\alpha + \beta X)}{1 + \exp(\alpha + \beta X)}$$

avec α et β les termes du modèle que l'on cherche à estimer, α étant la constante et β le coefficient associé à la variable X.

Notons que le choix de cette fonction permet de lier directement le paramètre β du modèle à l'odds ratio (OR) ou « rapport des cotes » qui est une mesure de la force d'association entre l'exposition à un facteur (respectivement $X = 1$ ou $X = 0$ s'il y a exposition ou non) et une maladie. Dans ce cas, OR = $(p_1 / 1 - p_1) / (p_0 / 1 - p_0)$ où p_1 et p_0 sont les probabilités de survenue de la maladie pour les sujets exposés et non exposés. L'OR est en fait le facteur par lequel est multiplié le risque de développer la maladie sous l'effet de l'exposition. Le modèle logistique est souvent écrit en utilisant une fonction Logit afin de rendre linéaire l'expression ; ainsi, on a : Logit $P(Y = 1|X) = Ln((P(Y = 1|X)/(1 - P(Y = 1|X)) = \alpha + \beta X$.

Finalement, OR = $exp(\beta)$, on en déduit que si OR est égal à 1 ($\Leftrightarrow \beta = 0$) alors l'association entre la maladie et le facteur étudié est nulle alors que OR > 1 ($\Leftrightarrow \beta \neq 0$) indique un facteur aggravant et OR < 1 ($\Leftrightarrow \beta \neq 0$) indique un facteur protecteur.

Le modèle logistique multiple

Le modèle précédent se généralise en présence de plusieurs variables explicatives,

$$P(Y = 1 \mid X_1,\dots,X_p) = \frac{\exp\ (\alpha + \sum_{i=1}^{p} \beta_i X_i)}{1 + \exp\ (\alpha + \sum_{i=1}^{p} \beta_i X_i)}$$

et

$$Logit\left(P(Y = 1 \mid X_1,\dots,X_p\right) = \alpha + \sum_{i=1}^{p} \beta_i X_i .$$

L'estimation des coefficients α et β_i (et donc des OR ajustés correspondants) du modèle se fait généralement par la méthode du maximum de vraisemblance *(maximum likelihood)* à partir d'un échantillon de n observations indépendantes et de même distribution (y_i, x_i). Cette approche vise à fournir une estimation des paramètres qui maximise la probabilité d'obtenir les valeurs réellement observées sur l'échantillon. Soit :

$$L(\alpha, \beta_i) = \prod_{j=1}^{n} p(x_j)^{yj} (1 - p(x_j))^{(1-yj)}$$

la fonction de vraisemblance qui s'utilise sous la forme logarithmique :

$$Log(L(\alpha, \beta_i)) = \sum_{j=1}^{n} y_j\, Log(p(x_j)) + (1 - y_j) Log(1 - p(x_j)).$$

l'estimation des α et β_i est alors réalisée après dérivation de l'expression de la log-vraisemblance en fonction de α et des β_i grâce à une méthode d'approximation numérique itérative : l'algorithme de Newton-Raphson.

La qualité d'ajustement du modèle ainsi obtenu peut alors être évalué au moyen des critères d'adéquation classiques : la statistique de vraisemblance ou déviance ($= -2\ Log\ (L(\alpha,\ \beta_i))$) et la statistique de Pearson ou selon les critères d'information de Akaike ($AIC = -2Log(L(\alpha,\beta_i)) + 2p$) et de Schwarz *(Bayesian Information Criterion)* ($BIC = -2Log(L(\alpha,\beta_i)) + pLog(n)$) qui pénalise la vraisemblance quand le nombre de paramètres augmentent. Le modèle est d'autant plus intéressant que la valeur de ces deux critères est faible.

Enfin, le principe de la sélection de variables (en d'autres termes le test de l'apport d'une variable ou d'un groupe de variables explicatives dans l'ajustement du modèle) est fondé sur la comparaison du modèle complet avec le modèle dont on a exclu la variable (ou les variables) à évaluer. Dans cette optique, les deux critères habituellement utilisés sont le test du rapport de vraisemblance et le test de Wald.

Pour illustrer la méthode, nous allons modéliser l'association entre les 5 auto-antigènes cités précédemment et la maladie définie comme la survenue d'un carcinome infiltrant. Le codage de la variable Y à expliquer est le suivant : 0 = population témoin (N = 93), 1 = population atteinte d'un carcinome infiltrant (N = 60).

L'analyse de l'exemple est réalisée avec le logiciel Stata 11.0 (StatCorp, College Station, TX) au moyen des fonctions *logit* et *logistic*, qui donne les résultats respectivement sous forme de coefficients ou directement en termes d'odds ratio. La commande *stepwise* permet de lancer des procédures automatiques (ascendante ou descendante) dont le principal avantage est l'examen systématique de l'ensemble des variables mais qui sont fondées uniquement sur des critères statistiques et qui peuvent ignorer de potentiels facteurs de confusion. Nous recommandons donc l'utilisation d'une procédure pas à pas descendante dont le principe est d'inclure dans le modèle initial toutes les variables et de retirer progressivement celles qui s'avèrent non significatives [2]. Le test du rapport de vraisemblance *(lrtest)* ou le test de Wald (*test* ou *testparm*) peuvent être utilisés à cet effet. Notons qu'il est nécessaire de stocker les résultats des modèles à comparer dans une macro variable à l'aide de la commande *estimates store* avant d'utiliser *lrtest*. Il est également important de vérifier que le rôle de confusion des variables enlevées soit négligeable, c'est-à-dire que les odds ratio ne varient pas de plus de 20 %, et finalement l'absence d'interactions dans le modèle final.

Dans notre exemple, seuls les auto-anticorps HSP60 et PPIA permettent d'expliquer la survenue d'un carcinome infiltrant. À titre d'exemple, le *tableau I* présente les valeurs du logarithme de la vraisemblance (LL pour *log-likelihood*), de AIC et de BIC pour certains modèles.

Le modèle logistique final est le suivant : $Logit\ P = -2{,}91 + 3{,}77 \times HSP60 + 4{,}25 \times PPIA$ *(figure 2)*. Dans la colonne « Coef », on trouve les valeurs estimés des coefficients α (_cons) et les β associés à chaque variable explicative du modèle. On trouve également la valeur de la statistique du rapport de vraisemblance :

$LR = -2 \times [LL(null) - LL(modèle)] = -2 \times [-102{,}4646 - (-81{,}09109)] = 49{,}75.$

Tableau I. Comparaisons des critères LL, AIC et BIC pour chaque modèle.			
Modèle	**LL**	**AIC**	**BIC**
null	-102,4646	206,9292	209,9596
FKBP52	-92,67531	189,3506	195,4115
HSP60	-93,8971	191,7942	197,8551
MUC1	-92,56502	189,13	195,1909
PPIA	-88,17678	180,3536	186,4144
PRDX2	-90,77467	185,5493	191,6102
HSP60 et PPIA	-81,09109	168,1822	177,2735

LL : log-vraisemblance ; AIC : critères d'information de Akaike ; BIC : *Bayesian Information Criterion.*

```
logit groupe hsp60 ppia
Logistic regression                               Number of obs   =        153
                                                  LR chi2(2)      =      42.75
                                                  Prob > chi2     =     0.0000
Log likelihood =  -81.09109                       Pseudo R2       =     0.2086
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hsp60 |   3.766036   1.376315     2.74   0.006     1.068509    6.463563
        ppia |   4.253858   1.138812     3.74   0.000     2.021828    6.485888
       _cons |  -2.914561   .5539275    -5.26   0.000    -4.000239   -1.828883
------------------------------------------------------------------------------
```

Figure 2. Sortie numérique associée à l'utilisation de la commande *logit* – modèle n° 1.

La dernière étape vise à vérifier l'adéquation du modèle aux données : en particulier, même si des variables s'avèrent significatives, le modèle peut ne pas bien expliquer la survenue de l'événement d'intérêt. Les commandes *estat gof* et *estimates stat* permettent d'obtenir directement les principaux critères d'adéquation : la statistique de Pearson et la déviance, ainsi que les critères AIC et BIC.

Le premier modèle déterminé est fondé sur les mesures quantitatives des auto-anticorps, il est également possible d'utiliser la régression logistique après catégorisation des variables. L'interprétation en termes d'odds ratio est alors plus adéquate. La commande à utiliser est identique pour des variables binaires ; en revanche, une variable qualitative nominale ne doit jamais être incluse telle quelle dans le modèle, car les estimations des paramètres dépendent du codage des variables. Il est nécessaire de la décomposer en variables indicatrices d'appartenance à une catégorie. Par exemple, pour une variable X à trois catégories, codée 1, 2 et 3, il est nécessaire de générer deux variables X2 et X3 qui prennent la valeur 1 si X = 2 et X = 3 respectivement, la catégorie 1 servant de catégorie de référence. La construction des variables indicatrices sous Stata

```
estat gof, group(10) table
Logistic model for groupe, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.1726 |     5 |   2.5 |    11 |  13.5 |    16 |
  |     2 | 0.2016 |     4 |   2.8 |    11 |  12.2 |    15 |
  |     3 | 0.2282 |     4 |   3.2 |    11 |  11.8 |    15 |
  |     4 | 0.2543 |     2 |   3.8 |    14 |  12.2 |    16 |
  |     5 | 0.2893 |     4 |   4.2 |    11 |  10.8 |    15 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.3423 |     4 |   4.7 |    11 |  10.3 |    15 |
  |     7 | 0.4318 |     4 |   6.2 |    12 |   9.8 |    16 |
  |     8 | 0.5603 |     6 |   7.5 |     9 |   7.5 |    15 |
  |     9 | 0.8244 |    12 |  10.6 |     3 |   4.4 |    15 |
  |    10 | 0.9998 |    15 |  14.4 |     0 |   0.6 |    15 |
  +--------------------------------------------------------+
       number of observations =       153
 number of covariate patterns =       153
            Pearson chi2(150) =       147.77
                  Prob > chi2 =         0.5362
```

Figure 3. Sortie numérique associée à l'utilisation de la commande *estat gof*.

```
estimates stat
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
       model |    153   -102.4646   -81.09109      3     168.1822    177.2735
-----------------------------------------------------------------------------
```

Figure 4. Sortie numérique associée à l'utilisation de la commande *estimates stat*.

est automatique avec la commande *xi*. Dans le cas de variables qualitatives ordinales, il est possible soit d'utiliser des indicatrices, soit d'inclure directement la variable dans le modèle afin de modéliser la relation dose-effet. Le calcul de l'odds ratio d'une classe particulière sera alors déterminé au moyen de la commande *lincom*.

Le modèle logistique polytomique

Le modèle logistique polytomique est une extension directe du modèle logistique pour réponse binaire au cas des réponses catégorielles. Notons k les différentes catégories avec $k = 1, \ldots, K$. L'association entre les 5 auto-antigènes et la maladie, définie à présent comme une variable réponse à trois modalités : 0 = population témoin (N = 93), 1 = population atteinte d'un CIS

(N = 82) et 2 = population atteinte d'un carcinome infiltrant (N = 60), est modélisée à l'aide de la commande Stata *mlogit* qui implémente la régression logistique polytomique et qui prend comme catégorie de référence la première modalité.

L'ordre des modalités de la réponse n'est pas pris en compte, on considère seulement une catégorie de référence et on modélise la probabilité de survenue de chacune des autres catégories par rapport à cette catégorie de référence. Dans ce contexte, le Logit $P(Y = 1 \mid X)$ est remplacé par le Logit généralisé calculé, par exemple, par rapport à la catégorie K, et c'est donc cette expression qui est écrite en fonction des variables explicatives X_i $(i = 1, \ldots, p)$:

$$Log \frac{P(Y = k \mid X)}{P(Y = K \mid X)} = \alpha_k + \sum_{i=1}^{p} \beta_{ik} X_i .$$

La probabilité d'appartenance d'un individu à la catégorie k est alors notée :

$$P(Y = k \mid X) = \frac{\exp\left(\alpha_k + \sum_{i=1}^{p} \beta_{ik} X_i\right)}{1 + \exp\left(\alpha_k + \sum_{k=1}^{K-1} \sum_{i=1}^{p} \beta_{ik} X_i\right)} \text{ si } k = 1, \ldots, K\text{-}1 \text{ et}$$

$$P(Y = K \mid X) = 1 - \sum_{k=1}^{K-1} P(Y = k \mid X),$$

et la log-vraisemblance à maximiser est la suivante :

$$Log(L(\alpha_k, \beta_{ik})) = \sum_{j=1}^{n} y_1 Log(p_1(x_j)) + \ldots + y_K Log(p_K(x_j)).$$

Le test de Wald permet de vérifier la significativité de chaque coefficient dans chaque régression et de réaliser également, comme le test du rapport de vraisemblance, une évaluation globale des variables introduites dans le modèle.

Seul l'auto-anticorps PRXD2 n'est pas conservé dans le modèle et les probabilités d'appartenance aux classes 1 (CIS) et 2 (cancer infiltrant) sont les suivantes :

$$Log \frac{P(Y = 1 \mid X = x)}{P(Y = 0 \mid X = x)}$$

$$= -2{,}13 + 1{,}69 \times \text{HSP60} + 4{,}69 \times \text{FKBP52} - 6{,}28 \times \text{MUC1} + 5{,}46 \times \text{PPIA}$$

$$Log \frac{P(Y=2 \mid X=x)}{P(Y=0 \mid X=x)}$$

$$= -2{,}71 + 2{,}38 \times HSP60 + 1{,}75 \times FKBP52 - 1{,}00 \times MUC1 + 3{,}75 \times PPIA$$

L'interprétation des coefficients est plus difficile que dans le cas dichotomique car ils sont relatifs à la catégorie de référence. L'explication du rapport des risques relatifs *(relative-risk ratio)*, c'est-à-dire comment une variable modifie le rapport de la probabilité étudiée sur la probabilité de base, est plus aisée à interpréter. Par exemple, considérons l'auto-antigène HSP60, une augmentation d'une unité du marqueur multiplie par 5 (≈ exp(1,693887)) la probabilité d'être atteint d'un cancer *in situ* et par 11 (≈ exp(2,384764)) la probabilité de développer un carcinome infiltrant. Ainsi, on peut conclure à l'implication de l'auto-antigène HSP60 à la fois dans la détection des cancers de stades précoces (CIS) et des carcinomes infiltrants, et présumer de la corrélation de HSP60 avec le grade de la maladie. Notons que pour obtenir directement le *relatif-risk ratio* à la place des coefficients, il est nécessaire d'utiliser l'option *rrr* avec la commande *mlogit*.

```
mlogit groupe hsp60 ppia

Multinomial logistic regression                   Number of obs   =        235
                                                  LR chi2(8)      =      81.69
                                                  Prob > chi2     =     0.0000
Log likelihood =  -213.6154                       Pseudo R2       =     0.1605
------------------------------------------------------------------------------
         grp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Contr_le     |  (base outcome)
-------------+----------------------------------------------------------------
CIS          |
       hsp60 |   1.693887   .8921945     1.90   0.058    -.0547823    3.442556
      fkbp52 |   4.698969   1.137986     4.13   0.000     2.468558     6.92938
        muc1 |  -6.282643   1.570671    -4.00   0.000    -9.361102   -3.204184
        ppia |   5.459719   1.605408     3.40   0.001     2.313178     8.60626
       _cons |  -2.130752   .4514883    -4.72   0.000    -3.015653   -1.245851
-------------+----------------------------------------------------------------
Infiltrant   |
       hsp60 |   2.384764   .8877937     2.69   0.007     .6447203    4.124808
      fkbp52 |   1.753917    1.13997     1.54   0.124     -.480384    3.988217
        muc1 |  -1.006432   1.471733    -0.68   0.494    -3.890976    1.878112
        ppia |   3.751221   1.600733     2.34   0.019     .6138408    6.888601
       _cons |  -2.709418   .4693228    -5.77   0.000    -3.629274   -1.789563
------------------------------------------------------------------------------
```

Figure 5. Sortie numérique associée à l'utilisation de la commande *mlogit* – modèle n° 2.

Le modèle logistique polytomique constitue la base de plusieurs autres modèles dans lesquels l'ordre des modalités de la réponse est pris en compte (les k catégories sont à présent ordonnées). Le modèle le plus simple et le plus utilisé, disponible sous Stata sous la commande *ologit*, est le modèle à odds proportionnels pour lequel l'ordre des modalités de la réponse est pris en compte mais l'effet d'une covariable est supposé le même pour le passage d'un niveau de réponse au niveau suivant quels que soient ces niveaux :

$$Log \frac{P(Y \geq k \mid X)}{P(Y < k \mid X)} = \alpha_k + \sum_{i=1}^{p} \beta_i X_i$$

où seule la constante dépend de k. Les probabilités s'écrivent alors :

$$P(Y \geq k \mid X) = \frac{\exp\left(\alpha_k + \sum_{i=1}^{p} \beta_i X_i\right)}{1 + \exp\left(\alpha_k + \sum_{i=1}^{p} \beta_i X_i\right)} \text{ si } k = 1, \ldots, K\text{-}1.$$

Enfin, citons le modèle logistique ordonné généralisé qui représente la généralisation la plus large du modèle proportionnel et pour lequel l'effet de toutes les variables explicatives varie selon le point de coupure sur les catégories de la variable dépendante.

Les courbes ROC

La précision d'un nouveau test diagnostique, par exemple le dosage d'un panel d'auto-antigènes à visée diagnostique, doit être comparée au test de référence *(Gold standard)* avant toute utilisation en routine comme outil diagnostique. C'est ainsi qu'il sera possible de connaître la validité (sensibilité/spécificité) de ce panel de dosage. Dans ce contexte, le résultat de l'examen biologique, noté T, associé au test diagnostique peut être dichotomique, c'est-à-dire soit positif (T+) si le résultat est supérieur ou égal au seuil de décision, soit négatif (T-) si le résultat est inférieur au seuil. Un tableau de contingence 2x2 comparant les deux populations malades (M) ou non malades (M-) résume alors le résultat de l'évaluation du test diagnostique. La sensibilité, la spécificité, les rapports de vraisemblance et les valeurs prédictives sont les indices les plus connus pour évaluer le pouvoir séparateur (« discriminant ») d'un test diagnostique, c'est-à-dire distinguer les sujets atteints d'une maladie de ceux qui ne le sont pas (il est fait référence ici aux qualités intrinsèques comme extrinsèques du test). En général, le seuil choisi, noté δ, ne permet pas de séparer complètement les deux populations. Le meilleur seuil que l'on puisse espérer est celui qui sépare entièrement la population malade (100 % sensible) de la population non malade (100 % spécifique), mais, en pratique, cette situation est utopique *(figure 6)*.

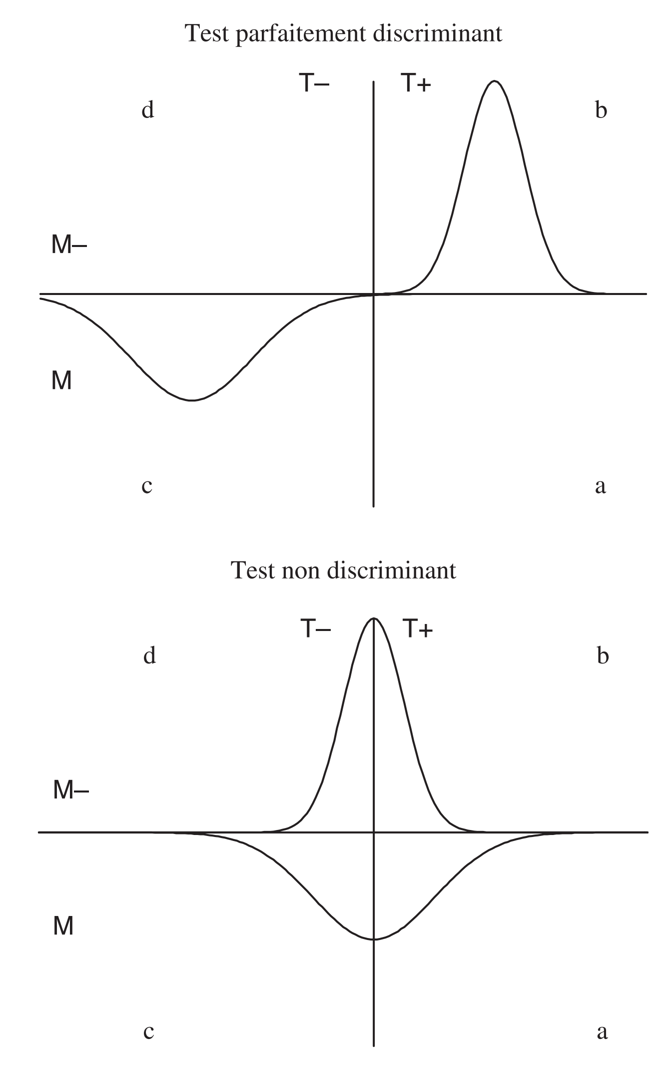


Figure 6. Qualité prédictive d'un test diagnostique.

Le résultat de l'examen biologique peut être également une variable quantitative continue (exemple : taux de PSA pour le cancer de la prostate) et la précision est alors mesurée par la méthode des courbes ROC.

La courbe ROC est la représentation graphique de la relation réciproque entre les taux de vrais et faux positifs, calculée pour toutes les valeurs seuils δ possibles. La sensibilité (Se = taux de vrais positifs) est la probabilité pour un sujet véritablement malade d'être correctement classé et 1 moins la spécificité (1 - Sp = taux de faux positifs) est la probabilité pour un sujet non malade d'être par erreur classé comme malade. Soit X et Y les résultats de ce même examen pour les

n_X individus de la population non malade et n_Y individus de la population malade respectivement ; X et Y sont indépendants. La courbe ROC est alors complètement définie par une fonction monotone croissante dans l'intervalle [0 ; 1] :

$$\{1 - F_x(\delta), 1 - F_y(\delta), \ \delta \in (-\infty, +\infty)\}$$

avec F_x et F_y les fonctions de répartition des distributions de X et Y respectivement. La courbe ROC peut également s'écrire comme une fonction de t dont l'expression est la suivante :

$$\mathrm{ROC}(t) = 1 - F_y\left(F_x^{-1}(1-t)\right) \ \textit{versus} \ t, \ t \in [0{,}1].$$

Les deux principales propriétés de la courbe ROC établie pour un test donné sont, d'une part, de proposer une estimation globale de la valeur prédictive d'un marqueur. En effet, la capacité prédictive entre deux populations est déterminée par la quantité de recouvrement entre les distributions des populations étudiées ; on peut éventuellement visualiser ce recouvrement avec les boxplots *(figure 1)*. Le degré de recouvrement (ou séparation des valeurs entre les groupes) détermine la forme et la position de la courbe ROC par rapport à la bissectrice (*cf.* exemple *figure 8*). Ce recouvrement est d'autant plus faible ou bien la discrimination est d'autant meilleure que la courbe ROC se rapproche de l'angle en haut à gauche. Ce point, caractérisé par une sensibilité et une spécificité égales à 1, est défini comme le point idéal à atteindre. D'autre part, la courbe ROC permet également de choisir le seuil optimal adapté à un but diagnostique précis ; la détermination de ce seuil est réalisée grâce à l'indice *J* de Youden (J = Se + Sp – 1) qui sera maximal en le point le plus discriminant.

L'aire sous la courbe ROC, notée AUC *(Area Under Curve)*, définie par $AUC = \int_0^t ROC(t)dt$ est l'indice le plus utilisé pour donner une estimation du pouvoir discriminant global du test. Sa valeur peut varier de 0,5 si le test n'est pas informatif à 1 si le test est parfaitement discriminant. Plusieurs estimations fondées sur des méthodes paramétrique, non paramétrique et semi-paramétrique ont été proposées pour l'estimation de la courbe ROC et de l'AUC. Dans la suite, nous allons nous restreindre à décrire les deux méthodes les plus utilisées et implémentées sous le logiciel Stata au travers des commandes *rocfit* et *roctab*.

Le modèle binormal

Le modèle binormal [3] suppose l'existence d'une fonction monotone permettant de transformer à la fois les distributions des populations malade et non malade en distributions normales. Une courbe ROC générée par des distributions qui respecte ce critère est complément spécifiée par deux paramètres $\alpha = (\mu_Y - \mu_X)/\sigma_Y$ et $\beta = \sigma_X/\sigma_Y$. L'expression de la courbe ROC est alors la suivante : $\mathrm{ROC_B}(t) = \Phi(\alpha + \beta\Phi^{-1}(t))$ où α est l'ordonnée à l'origine et β la pente de la courbe ROC binormale, Φ étant la fonction de répartition d'une loi normale centrée réduite. L'*AUC* a alors une forme analytique particulièrement simple :

$$AUC_B = \Phi\left(\frac{\alpha}{\sqrt{1+\beta^2}}\right).$$

Le modèle empirique

La méthode empirique [4] pour estimer une courbe ROC consiste, elle, à remplacer les fonctions de répartition F_X et F_Y par leurs estimations empiriques respectives dans l'expression : $\{1 - F_X(\delta), 1 - F_Y(\delta), \delta \in (-\infty, +\infty)\}$:

$$\hat{F}_Z(\delta) = \frac{Card\{Z_i \leq \delta\}}{n_Z} = \frac{\sum_{i=1}^{n_Z} I[Z_i \leq \delta]}{n_Z}.$$

avec $Z \in \{X, Y\}$. L'estimation de l'*AUC* sous la courbe ROC empirique est alors fondée sur la statistique U de Mann-Whitney :

$$AUC_E = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \left\{ I[Y_j > X_i] + \frac{1}{2} I[Y_j = X_i] \right\}.$$

Application

Pour illustrer ces méthodes, nous allons vérifier si l'antigène PPIA permet de discriminer la population de patientes atteintes d'un cancer infiltrant de la population témoin. La commande *rocfit* permet d'estimer les paramètres α et β du modèle binormal grâce à la méthode du maximum de vraisemblance ; il est au préalable nécessaire de normaliser les distributions par passage au logarithme, par exemple, et d'utiliser en suivant la commande *rocplot* afin de visualiser la courbe ROC. La commande *rocfit* ne permet pas de traiter plusieurs variables simultanément afin de pouvoir comparer les *AUC* respectifs *(figure 7)*. Un moyen pour pallier ce problème est d'utiliser les commandes *roctab*, *roccomp* et *rocgold* qui implémentent la méthode empirique et les tests associés. En particulier, *roccomp* implémente un test pour tester l'égalité de l'aire sous les courbes ROC obtenues en utilisant un algorithme suggéré par DeLong *et al.* [5]. L'antigène PPIA ne permet pas de discriminer de manière efficace la population des cas de celles des témoins ; en effet, l'aire sous la courbe ROC est faible quelle que soit la méthode d'estimation utilisée *(figure 8)* :

AUC_B = 0,63 (IC95 % [0,58 ; 0,77]) et AUC_E = 0,64 (IC 95 % [0,54 ; 0,75]) *(figure 9).*

Il en est de même pour l'antigène HSP60 : AUC_E = 0,60 (IC 95 % [0,51 ; 0,70]). Pour conclure, l'*AUC* entre ces deux marqueurs est également similaire (p = 0,57) *(figure 10).*

```
rocfit groupe log_ppia , continuous(10)

Binormal model of h_t1n0 on lppia                Number of obs   =        153
Goodness-of-fit chi2(5) =        8.88
Prob > chi2             =      0.1141
Log likelihood          =  -234.84866
------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   intercept |   0.509322   0.152279     3.34   0.001     0.210861    0.807783
   slope (*) |   0.458789   0.077788    -6.96   0.000     0.306327    0.611251
-------------+----------------------------------------------------------------
       /cut1 |  -2.008228   0.256736    -7.82   0.000    -2.511421   -1.505034
       /cut2 |  -0.478055   0.133132    -3.59   0.000    -0.738988   -0.217122
       /cut3 |   0.680137   0.133918     5.08   0.000     0.417662    0.942611
       /cut4 |   1.712584   0.221292     7.74   0.000     1.278859    2.146305
       /cut5 |   3.059553   0.510761     5.99   0.000     2.058480    4.060627
       /cut6 |   4.212546   0.770656     5.47   0.000     2.702088    5.723004
       /cut7 |   5.180296   1.025972     5.05   0.000     3.169427    7.191165
------------------------------------------------------------------------------
             |                  Indices from binormal fit
       Index |   Estimate   Std. Err.                     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    ROC area |   0.678292   0.049129                      0.582001    0.774582
    delta(m) |   1.110144   0.356798                      0.410834    1.809455
        d(e) |   0.698281   0.206633                      0.293288    1.103273
        d(a) |   0.654677   0.193855                      0.274729    1.034625
------------------------------------------------------------------------------
```

Figure 7. Sortie numérique associée à l'utilisation de la commande *rocfit.*

Analyse multivariée

Les méthodes décrites jusqu'à présent considèrent chaque marqueur (variables explicatives) de manière indépendante. Dans la plupart des études, on dispose de plusieurs marqueurs, et des méthodes multivariées sont intéressantes dans ce contexte, car elles peuvent permettre d'améliorer la précision du test diagnostique.

Le critère ROC généralisé est une technique qui est fondée sur la fonction discriminante de Fisher pour trouver une combinaison linéaire de marqueurs optimale dans le sens où l'aire sous la courbe ROC généralisée est maximale [6]. Plusieurs modèles ont été proposés pour l'estimation de l'aire sous la courbe ROC et des intervalles de confiance correspondants fondés :
– soit sur une estimation paramétrique pour le modèle généralisé : les distributions des variables sont supposées gaussiennes multivariées ;
– soit sur une estimation non paramétrique pour le modèle empirique.

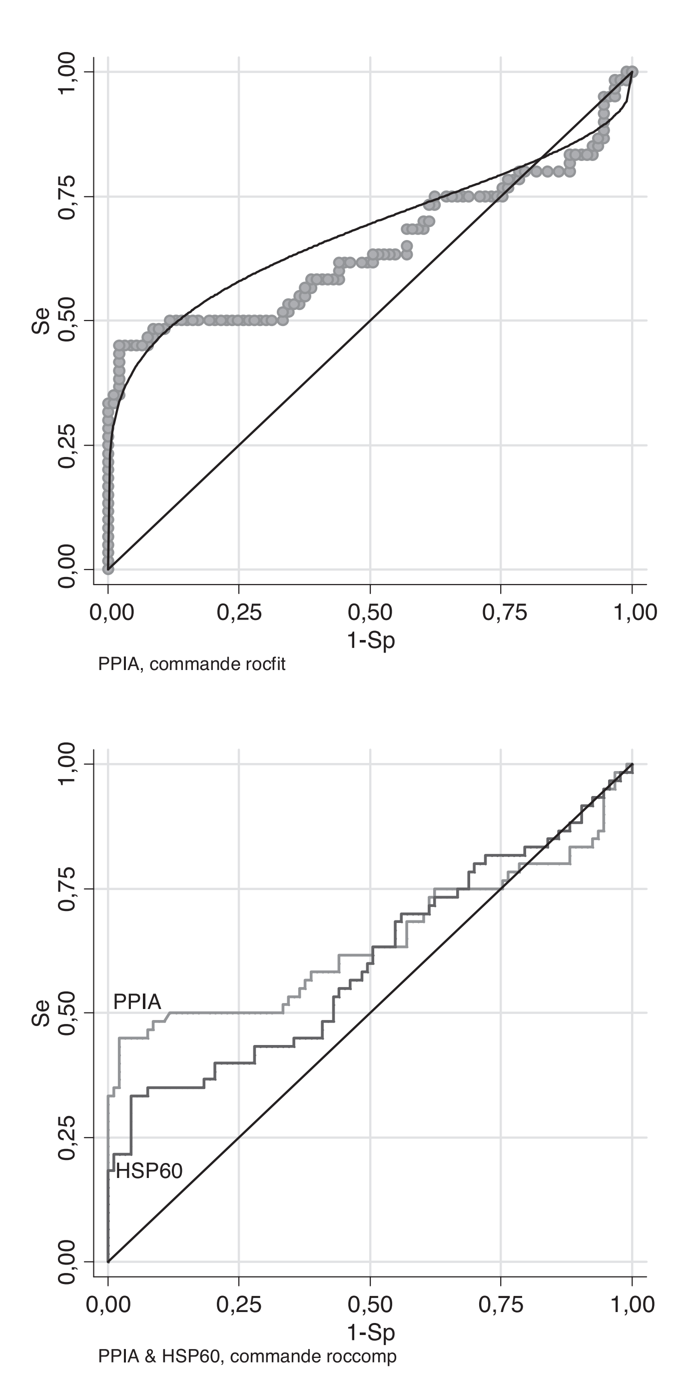


Figure 8. Courbes ROC pour les antigènes PPIA et HSP60.
Se : sensibilité ; 1-Sp : 1-spécificité.

```
roctab groupe ppia

                       ROC                    -Asymptotic Normal--
           Obs        Area     Std. Err.      [95% Conf. Interval]
          ------------------------------------------------------------
           153      0.6442       0.0518        0.54271      0.74564
```

Figure 9. Sortie numérique associée à l'utilisation de la commande *roctab*.

```
roccomp groupe ppia hsp60
                              ROC                    -Asymptotic Normal--
                   Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
ppia               153     0.6442       0.0518        0.54271     0.74564
hsp60              153     0.6034       0.0494        0.50653     0.70028
-------------------------------------------------------------------------
Ho: area(ppia) = area(hsp60)
    chi2(1) =      0.32       Prob>chi2 =   0.5721
```

Figure 10. Sortie numérique associée à l'utilisation de la commande *roccomp*.

Le logiciel mROC a été développé pour appliquer la méthode ci-dessus [7]. Dans notre exemple, la combinaison des 5 auto-anticorps est optimale pour discriminer la population atteinte d'un carcinome infiltrant de la population témoin ; le modèle obtenu avec le critère ROC généralisé (estimation non paramétrique) est le suivant :

$$Z = 0{,}947 \times HSP60 + 0{,}090 \times MUC1 - 0{,}568 \times PRDX2 + 1{,}218 \times PPIA + 0{,}162 \times FKBP52$$

et une estimation de l'aire empirique sous la courbe ROC égale à $AUC_E = 0{,}73$ (IC 95 % [0,64 ; 0,81]). Notons que la combinaison des deux marqueurs HSP60 et PPIA pour discriminer les deux populations donne une AUC similaire à la combinaison des 5 marqueurs :

$$Z = 0{,}926 \times HSP60 + 1{,}071 \times PPIA \text{ avec } AUC_E = 0{,}729 \text{ (IC 95 \% [0,63 ; 0,81])}.$$

Des méthodes alternatives fondées sur des critères différents de la fonction discriminante de Fisher ont également été proposées pour obtenir une combinaison linéaire de marqueurs. Par exemple, McIntosh et Pepe [8] ont utilisé un « score de risque » défini comme la probabilité d'être malade, en considérant des données de multiples marqueurs, comme une fonction optimale dans le sens où la courbe ROC est maximisée à chaque point. Pepe *et al.* [9] ont également défini une estimation de l'aire sous la courbe ROC fondée sur un score de classification pour dériver une combinaison de marqueurs optimisés pour la classification ou la prédiction.

Conclusions

La régression logistique et les courbes ROC sont deux méthodes statistiques très proches. En effet, l'objectif est de produire un modèle permettant de prédire les valeurs prises par la variable à expliquer pour la régression logistique, tandis que les courbes ROC permettent d'évaluer la performance d'une variable binaire. En particulier, le critère ROC généralisé permet de construire une combinaison linéaire de plusieurs covariables de la même manière que cela peut être réalisé avec la régression logistique. La différence réside dans le critère utilisé pour optimiser le modèle : la probabilité de la survenue de l'événement pour la régression logistique ou la fonction discriminante de Fisher en ce qui concerne les courbes ROC.

Références

1. Desmetz C, Bascoul-Mollevi C, Rochaix P, *et al*. Identification of a new panel of serum autoantibodies associated with the presence of in situ carcinoma of the breast in younger women. *Clin Cancer Res* 2009 ; 15 (14) : 4733-41.

2. Hosmer DW, Lemeshow S. *Applied logistic regression*. New York: Wiley Series in Probability and Mathematical Statistics, 2000 : 91-142.

3. Hanley JA. The use of the binormal model for parametric ROC analysis of quantitative diagnostic tests. *Stat Med* 1996 ; 15 : 1575-85.

4. Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic curve. *J Math Psychol* 1975 ; 12 : 387-415.

5. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 1988 ; 44 : 837-45.

6. Kramar A, Faraggi D, Ychou M, *et al*. Critères ROC généralisés pour l'évaluation de plusieurs marqueurs tumoraux. *Rev Epidemiol Sante Publique* 1999 ; 47 : 376-83.

7. Kramar A, Faraggi D, Fortuné A, Reiser B. mROC: A computer program for combining tumour markers in predicting disease states. *Comput Methods Programs Biomed* 2001 ; 66 : 199-207.

8. McIntosh MW, Pepe, MS. Combining several screening tests: Optimality of the risk score. *Biometrics* 2002 ; 58 : 657-64.

9. Pepe MS, Cai T, Longton G. Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* 2006 ; 62 (1) : 221-9.

Modèle de Cox et index pronostique

I. Le Ray, F. Kwiatkowski, F. Bonnetain

L'estimation des probabilités de survie au cours du temps par la méthode de Kaplan-Meier et la comparaison entre deux ou plusieurs groupes par le test du log-rank sont utiles pour les analyses de données censurées (*cf.* chapitre III.1 « Données de survie », page 129). Cependant, elles ne peuvent intégrer qu'une seule variable d'intérêt et, de fait, ne permettent pas de faire des analyses multivariées (multiparamétriques). Une analyse univariée est en général suffisante pour un essai randomisé, puisque les potentiels facteurs confondants sont contrôlés par la randomisation. En revanche, dans les études rétrospectives ou prospectives non randomisées, il est important de prendre tous les facteurs pronostiques en compte, ce que l'on nomme aussi « réaliser un ajustement » sur ces facteurs.

Pour étudier l'impact conjoint de plusieurs covariables sur la survie, il est nécessaire de modéliser la fonction de risque. Plusieurs modèles paramétriques ont été proposés : Weibull, Gompertz-Makeham... mais en cancérologie, c'est le modèle semi-paramétrique de Cox qui est le plus utilisé aujourd'hui [1].

L'idée initiale est que le risque *instantané* de survenue de l'événement (décès ou progression par ex.) est l'information dont l'intérêt prime pour les cliniciens. Celui-ci est un *rapport d'incidence*, par opposition à la régression logistique, qui fournit un *rapport de cotes* (odds ratio), qui est un rapport de proportions (*cf.* chapitre IV.1 « Régression logistique et courbes ROC », page 197). Devant un patient, le clinicien aimerait connaître, par exemple, la probabilité de survie à 5 ans ou bien la survie médiane. Une modélisation du risque de décès au cours du temps permet d'obtenir ces estimations pour un patient donné en fonction de son profil (stade de la maladie, statut ganglionnaire, âge, valeur de dosages biologiques, etc.). Le modèle de Cox est souvent utilisé car il pose moins d'hypothèses sur la forme de la fonction de risque au cours du temps et il est donc considéré comme un modèle semi-paramétrique, par opposition aux modèles paramétriques, comme le modèle de Weibull.

Base de calcul

Principes de modélisation

Si *t* est l'instant de survenue de l'événement, nous nous intéressons à la fonction de risque *h(t)* :

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

avec T l'instant du décès.

Cette fonction *h(t)* représente le risque de décès « juste après » l'instant *t* d'un patient qui était vivant à l'instant *t*.

Cette fonction peut prendre une forme quelconque au cours du temps et elle peut dépendre des valeurs prises par les *p* covariables étudiées $z_1, z_2, ..., z_p$.

Cette fonction *h(t)* est modélisée comme le produit d'une fonction de risque $h_0(t)$ dépendant uniquement du temps et d'une fonction de risque *g(z)*, dépendant uniquement de la valeur des covariables (donc de l'individu) :

$$h(t, z) = h_0(t) g(z)$$

Dans le modèle de Cox, la fonction *g(z)* prend une forme exponentielle :

$$g(z) = \exp(\beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_p z_p)$$

Ainsi, le risque de décès *h(t,z)* dépend d'une part du temps *t* et, d'autre part, de la valeur *z* des covariables pour chaque individu. Compte tenu de la relation entre risque et survie, on obtient :

$$S(t, z) = \exp\left(-\int_0^t h(t|z)\right) dz = S_0(t)^{g(z)}$$

Avec $S_0(t)$ la courbe de survie de base pour $g(z) = 1$, ce qui correspond à la situation où toutes les covariables *z* sont initialisées à la valeur zéro.

Ainsi, le risque relatif (HR pour *hazard ratio*) de décès d'un individu par rapport à un autre s'obtient par le rapport des risques de chaque individu. Comme la fonction de base $h_0(t)$ est supposée la même pour chaque individu, le risque relatif $HR_{i,j}$ entre deux individus *i* et *j* est égal au rapport des valeurs des covariables de ces deux individus $g(z_i)/g(z_j)$:

$$HR_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{h_0(t) g(z_i)}{h_0(t) g(z_j)}$$

Un calcul simple permet dès lors d'exprimer le risque instantané de décès en fonction des valeurs des covariables $z_1, z_2, ..., z_p$. Par exemple, un patient qui prend comme valeurs de covariables le vecteur $(z_1, z_2, ..., z_p)$ = (1,0, ..., 0) par rapport à un patient qui prend comme valeurs de covariables le vecteur $(z_1, z_2, ..., z_p)$ = (0,0, ..., 0), aura un risque relatif égal à l'exponentielle du coefficient β_1.

$$HR_{z_1=1} = \frac{\exp(\beta_1(z_1=1))}{\exp(\beta_1(z_1=0))} = \exp(\beta_1)$$

En pratique, ces coefficients sont obtenus par la méthode dite du maximum de vraisemblance. Seule la partie de la vraisemblance faisant intervenir ces coefficients est retenue ; on parle de vraisemblance partielle ou vraisemblance de Cox. Trois tests, dont les résultats sont en général concordants, peuvent être utilisés pour tester l'hypothèse nulle $H_0 : \beta = 0$ *vs* $H_1 : \beta \neq 0$ et ainsi évaluer la significativité. Ce sont les tests de Wald, du score et du rapport de vraisemblance [2].

Si z_1 est une variable binaire, l'expression de HR signifie que le risque instantané de décès est multiplié par $exp(\beta_1)$ pour $z_1 = 1$ par rapport à $z_1 = 0$. Cette valeur est ajustée pour les autres covariables, puisqu'elles ont été inclues dans le modèle servant au calcul de β_1.

Dans le cas où z_1 n'est pas une variable binaire, l'expression précédente indique que, pour une augmentation de z_1 d'une unité (ou changement à une catégorie supérieure), le rapport de risque instantané est multiplié par $exp(\beta_1)$. On fait donc l'hypothèse que le fait de changer la valeur de z_1 d'une unité a le même effet, quelle que soit la valeur de z_1. Ainsi, le fait de passer de 0 à 1 est équivalent en termes de risque à passer de 1 à 2, de 2 à 3 ou de 28 à 29. Cette hypothèse est appelée hypothèse de log-linéarité : quand on additionne 1 à la valeur de z_1, on multiplie le risque instantané par $exp(\beta_1)$ (c'est-à-dire qu'on ajoute β_1 à son logarithme).

Une variable pour laquelle le coefficient β est négatif est associée à un rapport de risque inférieur à 1 : elle a un effet protecteur par rapport à l'événement considéré. À l'inverse, une variable dont le coefficient β est positif est associée à un rapport de risque supérieur à 1. Elle a un effet délétère. À titre d'exemple avec 200 sujets par groupe et en l'absence de perdus de vue avant la date de point, si on observe 80 et 60 décès dans les groupes A et B respectivement, les risques de décès sont de 0,40 et 0,30, ce qui correspond à un risque relatif de 0,30/0,40 = 0,75. La réduction de risque est de 25 % dans le groupe B par rapport au groupe A. Ainsi, un essai thérapeutique de supériorité comparant un nouveau traitement à un traitement de référence conclut à la supériorité du nouveau traitement si le *hazard ratio* est significativement inférieur à 1, montrant un bénéfice en termes de réduction de risque.

Choix du codage

Plusieurs considérations interviennent dans le choix du codage des variables. Les variables continues présentent plus d'avantages pour la modélisation que les variables découpées en classes, notamment en ce qui concerne la puissance et la précision des estimations [3] *(tableau I)*. Quant aux choix du découpage, la plupart des variables utilisées en cancérologie sont déjà présentées

Tableau I. Choix du codage des variables et conséquences.		
Item	**Codage catégoriel**	**Codage continu**
Puissance	Moins de puissance	Plus de puissance
Estimation	Perte de précision	Plus de précision
Hypothèse de log-linéarité	Constant par intervalle Discontinuité dans l'estimation du risque d'une catégorie à l'autre	Constant entre deux catégories
Découpage en catégories	Arbitraire Différent selon les études Intervalles de confiance larges pour les catégories extrêmes	Pas de découpage
Interprétation	Plus facile en termes d'odds ratio, etc.	Possible de comparer le risque entre deux valeurs éloignées

par catégorie, comme le T et le N, bien que ces classifications proviennent de mesures quantitatives à l'origine. La décision du choix du codage dépend également du degré de précision que l'on est prêt à accorder aux mesures brutes. Les autres considérations sont d'ordre pratique, car des protocoles de traitement sont adaptés selon ces classifications. À l'heure actuelle, il n'est pas envisageable d'adapter les doses de chimiothérapie selon la taille de la tumeur mesurée en millimètres.

Hypothèses sous-jacentes

Trois hypothèses sont requises pour construire ce modèle :

- **la censure est non informative** : c'est déjà le cas pour que les estimations des probabilités de survie selon la méthode de Kaplan-Meier soient non biaisées ;
- l'effet des variables est **log-linéaire** ;
- **les risques sont proportionnels** : le rapport de risque de l'effet des covariables est constant dans le temps.

Ces trois hypothèses doivent être vérifiées à chaque fois qu'un modèle de Cox est réalisé.

Censure non informative

L'hypothèse de la censure non informative est difficile à valider. En effet, par définition, il n'y a pas d'information sur les données censurées ; cependant, on doit s'assurer qu'il n'y a pas de biais majeur à ce niveau [4]. Dans un essai comparatif, on peut vérifier que la distribution des données

censurées est similaire entre les groupes. Par exemple, si on étudie le temps jusqu'à la perte d'autonomie chez le sujet âgé, il est nécessaire de s'assurer que les patients perdus de vue pour cause de déménagement ne sont pas placés en maison de retraite.

Log-linéarité

L'hypothèse de log-linéarité ne nécessite de vérification que pour les variables continues ou qualitatives ordonnées à plus de deux classes.

Chaque classe sera considérée comme une variable autonome, qui sera comparée à une classe de référence (découpage en tableau disjonctif complet ou *dummy variables*). Ainsi, on peut vérifier que, pour passer d'une classe à une autre, il suffit bien de multiplier les rapports de risque instantané. Si pour passer de la classe 1 à la classe 2, le rapport de risque instantané est de 3, alors pour passer de la classe 1 à la classe 3, le rapport de risque doit être de $3 \times 3 = 9$. Si ce n'est pas le cas, alors la variable ne peut être conservée sous forme ordinale dans le modèle, et on devra garder le découpage en tableau disjonctif complet dans le modèle final.

Pour les variables continues, on crée trois à cinq classes d'amplitudes égales, en utilisant préférentiellement des seuils reconnus, puis on procède de la même façon que pour les variables qualitatives ordinales. Si l'hypothèse de log-linéarité est vérifiée, la variable peut être conservée telle quelle dans le modèle final. Sinon, il est nécessaire de lui appliquer une transformation mathématique qui restaurera la log-linéarité ou alors de l'intégrer dans le modèle sous forme de classes en tableau disjonctif complet. On aura alors dans le modèle final un coefficient par classe de cette variable (sauf pour la classe de référence, dont le coefficient est 1 par définition).

Risques proportionnels

L'hypothèse des risques proportionnels doit d'abord être vérifiée graphiquement : les courbes de survie doivent suivre une évolution semblable dans les différents groupes et, en particulier, ne doivent pas se croiser. Cette hypothèse peut être testée grâce aux résidus de Schoënfeld [5]. Ces quantités mesurent l'apport supplémentaire d'information qui serait fourni par la prise en compte du temps dans la modélisation de risque instantané associé à une variable. Ainsi, si les résidus sont faibles (test non significatif), le fait de prendre le temps en compte n'apporterait qu'une information de faible importance. La *figure 1* présente les résidus de Schoënfeld pour la variable « année de diagnostic » de l'exemple présenté ci-dessous : le test associé est non significatif. L'hypothèse des risques proportionnels n'est pas rejetée.

Si, au contraire, le test sur les résidus est significatif, l'effet du temps ne peut être négligé.

Il existe des situations pour lesquelles l'hypothèse des risques proportionnels n'est pas vérifiée. Ainsi, dans le cadre d'un essai comparant une chimiothérapie seule à une intervention chirurgicale seule par exemple, l'acte chirurgical peut entraîner une augmentation du nombre de décès à court terme, mais peut être globalement bénéfique à moyen ou long terme. L'exemple présenté dans la *figure 2* montre des courbes de risque et de survie qui se croisent avec une mortalité

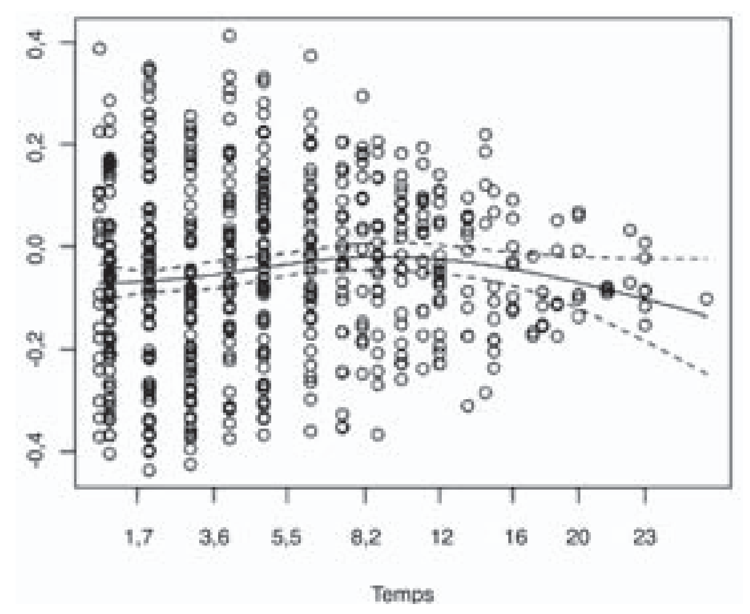


Figure 1. Résidus de Schoënfeld pour la variable « année de diagnostic ».

d'abord plus importante dans le groupe chirurgie (ligne solide), et le risque instantané qui diminue au cours du temps dans ce groupe alors qu'il augmente dans le groupe chimiothérapie (ligne pointillée). On peut remarquer que les fonctions de risque se croisent à un temps (3 unités) plus précoce que les courbes de survie (7 unités). Dans cette situation, il est nécessaire de prendre le temps en compte dans la modélisation.

Si ce phénomène concerne une variable d'ajustement, une solution consiste à utiliser un modèle de Cox stratifié sur cette variable et à évaluer la valeur pronostique des autres covariables. On peut donc tenir compte de cette variable dans l'analyse, mais sans exprimer son effet propre.

Il est également possible de garder la variable dans le modèle en intégrant le temps, de façon plus ou moins complexe selon les besoins du modèle. Dans l'exemple représenté sur la *figure 2*, le temps peut être divisé en deux périodes : avant et après 3 unités de temps. Une autre solution consiste à répartir l'axe de temps en deux périodes selon le nombre d'événements observés. On calcule alors un rapport de risque instantané pour chaque période. Il existe des manières plus complexes d'intégrer l'effet du temps, par exemple par une modélisation de l'interaction avec le risque instantané, sous forme polynomiale, le plus souvent de degré 3 [6].

Adéquation du modèle

Une façon de tester l'adéquation du modèle est de calculer le rapport de vraisemblance, qui doit être maximisé. Le principe est de calculer la probabilité d'observer exactement les données réelles étant donné le modèle et de la diviser par la probabilité d'observer ces mêmes données lors

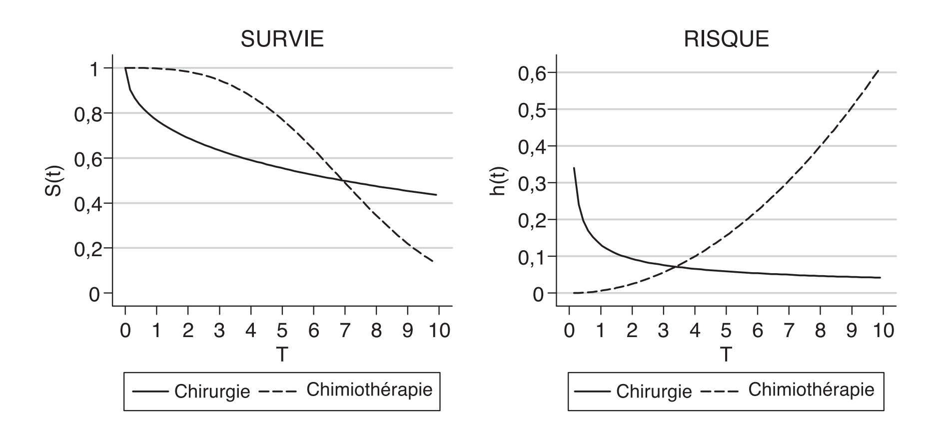


Figure 2. Exemple de courbes de survie et de risque qui se croisent.

d'un tirage aléatoire équiprobable. Par exemple, une pièce de monnaie jetée 10 fois tombe 7 fois sur pile et 3 fois sur face. On fait l'hypothèse que la probabilité de tomber sur face est de 0,7. La probabilité d'observer 7 fois face sur 10 tirages est de : $0{,}7^7 * 0{,}3^3 * C_7^3$. Si la probabilité d'obtenir face était de 0,5, on aurait comme résultat : $0{,}5^7 * 0{,}5^3 * C_7^3$. Le rapport de vraisemblance du modèle est donc de : $0{,}7^7 * 0{,}3^3 / 0{,}5^{10} = 2{,}27$. Plus le nombre d'observations est important et plus le modèle est adéquat, plus le rapport de vraisemblance augmente.

Exemple

L'exemple suivant présente une analyse des facteurs influençant la survie parmi des patientes résidant en Côte-d'Or atteintes d'un cancer du sein invasif opérable diagnostiqué entre 1982 et 2006. Pour réaliser cette analyse, il faut d'abord sélectionner les variables d'intérêt : les stades T (T1, T2, T3, T4) et N (N0, N+), l'âge de la patiente au moment du diagnostic (en années) et l'année du diagnostic. Elles ont été retenues sur la base de données de la littérature et sur des arguments cliniques. Dans cet exemple, il y avait 1 620 patientes, dont 533 sont décédées.

Analyse univariée

La première étape consiste à réaliser une analyse univariée pour chacune des quatre variables qui nous intéressent *(tableau II)*.

Lorsqu'on réalise une analyse univariée, les quatre variables apparaissent liées au pronostic. Trois variables ont un HR supérieur à 1, ce qui implique que le risque de décès augmente pour les valeurs croissantes des variables « stade T », « stade N » et « âge ». Le HR inférieur à 1 pour la

Tableau II. Analyse univariée.					
Variable (valeur au moment du diagnostic)	**Coefficient β**	**HR = *exp*(β)**	***p***	**RV**	**Critère AIC**
Stade T	0,52	1,68	< 0,0001	124,07	7 008,41
Stade N (référence : N0)	0,58	1,79	< 0,0001	38,23	7 094,26
Âge (années)	0,02	1,02	< 0,0001	22,36	7 110,12
Année de diagnostic	-0,06	0,94	< 0,0001	58,29	7 074,59

HR : *hazard ratio* ; RV : rapport de vraisemblance ; AIC : critère d'information de Akaike.

variable « année du diagnostic » implique que le risque de décès diminue pour les patientes diagnostiquées plus récemment. Ce choix de codage implique aussi que le risque de décès est constant entre des catégories adjacentes de même taille, car le HR dépend du système de codage.

Par exemple pour le stade T, le risque de décès est 1,68 pour un patient présentant un stade T2 par rapport à un patient au stade T1, il est aussi de 1,68 pour un patient présentant un stade T3 par rapport à un patient au stade T2, et ainsi de suite, car cette variable est codée 1, 2, 3 et 4.

L'étape suivante consiste à valider le codage continu car les interprétations des risques en dépendent. Par exemple, peut-on considérer le stade T, l'âge et l'année de diagnostic comme des variables continues ayant un effet log-linéaire sur la survie ? Pour ce faire, on scinde les variables continues (âge et année de diagnostic) en cinq classes d'amplitude égale. Chaque classe peut être analysée indépendamment des autres. L'interprétation des coefficients et du HR dépend de la catégorie de référence choisie. Dans cet exemple, la catégorie T2 de la variable « stade T » a été choisie comme catégorie de référence. Trois variables indicatrices définies comme appartenance ou non aux catégories T1, T3 et T4 sont ensuite introduits dans le modèle. Les résultats indiquent que les patients présentant un stade T1 ont un risque de décès de 0,47 par rapport aux patients au stade T2, ce qui se traduit par une réduction du risque de 53 %. En revanche, les patients présentant un stade T3 avec un HR de 1,44 ont une augmentation du risque de 44 % par rapport aux patients au stade T2 *(tableau III)*. En se basant sur les résultats du modèle avec la variable « stade T » codé de manière continue (HR = 1,68), l'estimation du risque est en augmentation de 68 % entre chaque catégorie, alors qu'il est de 113 %, 44 % et 53 % respectivement avec le modèle catégoriel.

Le critère d'information de Akaike (AIC = $-2Log(L(\alpha,\beta_i)) + 2p$), qui pénalise la vraisemblance quand le nombre de paramètres augmente, permet de comparer des modèles et ainsi choisir un codage qui reflète mieux les risques de décès entre les catégories *(tableau III)*. Le modèle est d'autant plus intéressant que la valeur de ce critère est faible.

Le critère AIC est plus faible pour le modèle qui prend en compte le stade T comme variable catégorielle, plutôt que comme variable continue (7 005,89 *vs* 7 008,41). De même, pour l'âge, le critère AIC passe de 7 110,12 (codage continu) à 7 107,45 (expression en classes). Pour ces deux variables, il est préférable de choisir le codage en classes plutôt que le codage en continu. En

Tableau III. Analyse univariée des variables « stade T », « âge » et « année » en catégories.

Variable (valeur au moment du diagnostic)		Coefficient β	HR = $exp(\beta)$	p	RV	Critère AIC
Stade	T1	-0,76	0,47	< 0,0001	130,59	7 005,89
	T2*	0	1			
	T3	0,36	1,44	0,0040		
	T4	0,79	2,20	< 0,0001		
Âge	< 40	-0,08	0,92	0,56	31,04	7 107,45
	40-49	-0,2	0,82	0,09		
	50-59*	0	1			
	60-69	0,29	1,33	0,02		
	≥ 70	0,59	1,81	0,0001		
Année	< 1987	0,41	1,51	< 0,0001	53,8	7 084,95
	1987-1991	0,24	1,27	0,06		
	1992-1996*	0	1			
	1997-2001	-0,30	0,74	0,03		
	> 2001	-0,87	0,42	< 0,0001		

* Catégorie de référence ; HR : *hazard ratio* ; RV : rapport de vraisemblance ; AIC : critère d'information de Akaike.

revanche, le critère AIC associé au modèle de Cox analysant l'effet de l'année de diagnostic est meilleur pour l'expression continue (7 074,59 *vs* 7 084,95). C'est donc ce codage qui sera utilisé par la suite.

Analyse multivariée

Les résultats du modèle multivarié sont présentés dans le *tableau IV*. Chacune des quatre variables reste associée de manière significative au risque de décès, montrant que :

- les risques de décès sont différents selon le stade T quels que soient le stade N, l'âge et l'année du diagnostic ;
- le risque est plus élevé pour les patients N+ quels que soient le stade T, l'âge et l'année de diagnostic ;
- les risques de décès sont différents selon l'âge quels que soient le stade T, le stade N et l'année du diagnostic ;
- les risques de décès sont différents selon l'année du diagnostic quels que soient le stade T, le stade N et l'âge.

Tableau IV. Analyse multivariée.					
Variable (valeur au moment du diagnostic)		**Variable du modèle**	**Coefficient β *associé***	**HR (IC 95 %)**	***p***
Stade T	T1	Z_{11}	-0,662	0,516 (0,42-0,63)	< 0,0001
	T2		0		
	T3	Z_{13}	0,277	1,320 (1,03-1,69)	0,03
	T4	Z_{14}	0,513	1,671 (1,23-2,27)	0,00
Stade N	N0		0	1	
	N1	Z_{21}	0,241	1,273 (1,05-1,55)	0,02
Âge	< 40	Z_{31}	0,043	1,043 (0,79-1,39)	0,77
	40-49	Z_{32}	-0,173	0,841 (0,66-1,07)	0,15
	50-59		0	1	
	60-69	Z_{34}	0,333	1,396 (1,10-1,78)	0,01
	≥ 70	Z_{35}	0,623	1,864 (1,37-2,54)	< 0,0001
Année de diagnostic		Z_{41}	-0,044	0,957 (0,94-0,97)	< 0,0001

HR : *hazard ratio*.

Le modèle multivarié s'écrit de la manière suivante :

$$h(t,z) = h_0(t)(g(z_1) + g(z_2) + g(z_3) + g(z_4))$$

Avec les variables z_1, z_2, z_3 et z_4 correspondant aux variables « stade T », « stade N », « âge » et « année de diagnostic » respectivement. En fonction du codage choisi pour chaque variable, les quatre fonctions *g(z)* ont des formes différentes. La variable z contient deux indices : le premier indice correspond à la variable et le deuxième indice correspond à la catégorie de cette variable. Par exemple, la variable z_{34} correspond à la catégorie 60-69 ans de la variable « âge ».
Pour les catégories de la variable « stade T », la fonction $g(z_1)$ prend la forme suivante :

$$g(z_1) = \exp(\beta_{11}z_{11} + \beta_{13}z_{13} + \beta_{14}z_{14})$$

Avec le codage suivant (stade T2 étant la catégorie de référence) :

$z_{11} = 1$, si stade T = 1 ; sinon $z_{11} = 0$

$z_{13} = 1$, si stade T = 3 ; sinon $z_{13} = 0$

$z_{14} = 1$, si stade T = 4 ; sinon $z_{14} = 0$

Pour la variable « stade N », la fonction $g(z_2)$ prend la forme suivante :

$$g(z_2) = \exp(\beta_{21}z_{21})$$

Avec le codage suivant pour les variables indicatrices :

$z_{21} = 1$, si stade N = 1 ; sinon $z_{21} = 0$

Pour les catégories de la variable « âge », la fonction $g(z_3)$ prend la forme suivante :

$$g(z_3) = \exp(\beta_{31}z_{31} + \beta_{32}z_{32} + \beta_{34}z_{34} + \beta_{35}z_{35})$$

Avec le codage suivant (âge ∈ {50 : 59} étant la catégorie de référence) :

$z_{31} = 1$, si âge < 40 ; sinon $z_{31} = 0$

$z_{32} = 1$, si âge ∈ {40 : 49} ; sinon $z_{32} = 0$

$z_{34} = 1$, si l'âge ∈ {60 : 69} ; sinon $z_{34} = 0$

$z_{35} = 1$, si l'âge ≥ 70 sinon $z_{35} = 0$

Il n'y a pas de variable z_{33} correspondant à la catégorie de référence : elle est en effet caractérisée par $z_{31} = 0$ *et* $z_{32} = 0$ *et* $z_{34} = 0$ *et* $z_{35} = 0$.

Pour la variable « année de diagnostic », la fonction $g(z_4)$ prend la forme suivante :

$$g(z_4) = \exp(\beta_{41}z_{41})$$

Avec le codage suivant :

z_{41} = année diagnostic ∈ {1982 : 2006} – 1995

$z_{41} \in \{-13 : 11\}$

Le fait de centrer le codage pour l'année de diagnostic permet de mieux situer le calcul des risques relatifs associés aux combinaisons des valeurs de covariables.

Le rapport de vraisemblance du modèle est de 196,1, plus élevé que pour n'importe laquelle des analyses univariées. Le critère AIC est égal à 6 952,38, montrant une meilleure adéquation du modèle.

À partir des estimations des 9 paramètres de ce modèle, il est possible d'obtenir des estimations des différentes combinaisons de ces variables relatives à la catégorie de référence, en identifiant les catégories impliquées et en associant le codage utilisé. Le risque relatif vaut 1 pour les patientes présentant un stade T2 ($z_{11} = 0$, $z_{13} = 0$, $z_{14} = 0$), N0 ($z_{21} = 0$), d'un âge compris entre 50 et 59 ans ($z_{31} = 0$, $z_{32} = 0$, $z_{34} = 0$, $z_{35} = 0$) et diagnostiquées en 1995 ($z_{41} = 0$). Toutes les variables valent zéro pour cette combinaison.

Par exemple, le risque de décès le plus faible correspond aux patientes présentant un stade T1 ($z_{11} = 1$, $z_{13} = 0$ $z_{14} = 0$), N0 ($z_{21} = 0$), âgées entre 40 et 49 ans ($z_{31} = 0$, $z_{32} = 1$, $z_{34} = 0$, $z_{35} = 0$) et diagnostiquées en 2006 ($z_{41} = 11$). Ce risque vaut 0,318, car il suffit de multiplier les coefficients β_{ij} associés à chaque catégorie par la valeur de la covariable, en faire la somme puis prendre l'exponentielle : $\exp(\beta_{11}*z_{11} + \beta_{41}*z_{41}) = \exp(-0{,}661*1 - 0{,}044*11)$.

Le risque de décès le plus élevé correspond aux patientes présentant un stade T4 ($z_{11} = 0$, $z_{13} = 0$ $z_{14} = 1$), N+ ($z_{21} = 1$), âgées de 70 ans ou plus ($z_{31} = 0$, $z_{32} = 0$, $z_{34} = 0$, $z_{35} = 1$) et diagnostiquées en 1982 ($z_{41} = -13$). Ce risque vaut 4,740, toujours en multipliant les coefficients β_{ij} associés à chaque catégorie par la valeur de la covariable, en faisant la somme puis en prenant l'exponentielle : $\exp(\beta_{14}*z_{14} + \beta_{21}*z_{21} + \beta_{36}*z_{36} + \beta_{41}*z_{41}) = \exp(0,513*1 + 0,241*1 + 0,623*1 - 0,044*(-13))$.

De cette manière, il est possible d'obtenir des estimations du risque relatif correspondant à chaque combinaison de covariables, ce qui permet de classer les patientes selon le risque de décès connaissant leurs caractéristiques. Cette technique permet l'obtention de groupes pronostiques de risque similaire, qui peuvent être ensuite utilisés pour appliquer des stratégies thérapeutiques adaptées au pronostic.

Conseils pratiques

Comme tout modèle statistique, les estimations des paramètres issues d'un modèle de Cox seront plus ou moins précises selon que l'on dispose d'un petit nombre de données ou bien d'un grand nombre. On parle ici de « stabilité » du modèle. En particulier, si le nombre de covariables est important alors que le nombre de patients inclus ne l'est pas, deux écueils peuvent se présenter :
– le calcul matriciel itératif peut ne pas « culminer » (converger) à une estimation du maximum de vraisemblance des paramètres ;
– le modèle peut culminer, mais les coefficients et les écarts-types associés ne sont pas très fiables. Par exemple, en ajoutant quelques données supplémentaires, on peut aboutir à des conclusions différentes. Par la suite, on devra s'interroger sur toute généralisation des conclusions tirées du modèle.

En pratique, il est conseillé d'avoir un minimum de 10 événements (décès, rechutes, etc.) par covariable incluse dans le modèle [7]. Ainsi, dans l'exemple ci-dessus, avec 9 paramètres introduits dans le modèle, un minimum de 90 décès serait nécessaire dans l'échantillon étudié pour garantir une bonne stabilité des estimations.

Conclusions

Le modèle de Cox permet l'étude multivariée de données censurées en se centrant sur le risque instantané de décès ou tout autre événement de survenue unique au cours du temps. Il est particulièrement bien adapté à l'évaluation de la valeur pronostique de plusieurs covariables, qu'elles soient binaires, quantitatives ou qualitatives ordinales. Il est primordial de vérifier *a posteriori* la validité des hypothèses du modèle. En cas de non-respect, le modèle doit être adapté pour tenir compte des spécificités des données. Les modèles peuvent devenir assez complexes, mais il faut être conscient qu'un modèle trop complexe, même s'il décrit bien le risque, ne sera pas utilisé.

Références

1. Cox DR. Regression models and life-tables (with discussion) *J Royal Stat Soc B* 1972 ; 34 : 138-220.

2. Hill C, Com-Nougué C, Kramar A, *et al*. *Analyse statistique des données de survie*. Paris : Flammarion, 1996.

3. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: A bad idea. *Stat Med* 2006 ; 25 : 127-41.

4. Timsit JF, Alberti C, Chevret S. Le modèle de Cox. *Rev Mal Respir* 2005 ; 22 (6 Pt 1) : 1058-64.

5. Schoenfeld D. Chi-squared goodness-of-fit tests for the proportional hazards regression model. *Biometrika*. 1980 ; 67 (1) : 145-53.

6. Huang JZ. Polynomial spline estimation and inference of proportional hazards regression models with flexible relative risk form. *Biometrics* 2006 ; 62 : 793-802.

7. Harrell FE, Lee KL, Mark DB. Tutorial in biostatistics: Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996 ; 15 : 361-87.

Modèle de Cox avec covariables dépendant du temps

J.M. Boher, B. Esterni

Le modèle de régression de Cox [1] est le modèle de référence pour estimer l'effet d'une ou de plusieurs variables sur des données de survie censurées dans le domaine de la recherche clinique en cancérologie. Les inférences sur les paramètres du modèle reposent sur une fonction de vraisemblance partielle et la proportionnalité des risques instantanés au cours du temps. Dans son article séminal [2], DR. Cox considère déjà le cas d'un modèle avec covariables dépendant du temps. Il faudra cependant attendre plusieurs années, et notamment les travaux de Aalen [2] pour établir les bonnes propriétés des estimateurs de maximum de vraisemblance partielle en présence de variables dépendant du temps [3]. Depuis, le modèle à risques proportionnels initié par Cox a été étendu à différentes situations diverses et variées. Pour une revue détaillée des applications nombreuses du modèle, le lecteur pourra se référer à l'ouvrage de Therneau et Grambsch [4].

Une situation souvent rencontrée en cancérologie concerne l'étude de l'impact d'un événement intermédiaire (la réponse à un traitement par ex.) sur la survie globale ou la survie sans maladie des patients. Nous aborderons dans ce chapitre les aspects pratiques de la modélisation à partir d'un exemple, l'étude de l'influence de la rechute sur la survie sans maladie d'une cohorte de patients atteints d'un cancer de la prostate.

Exemple : rechute et survie de patients atteints d'un cancer de la prostate

Entre septembre 1995 et décembre 2006, 750 patients ont été traités à l'Institut Paoli-Calmettes pour un cancer de la prostate de risque faible ou intermédiaire : 230 par irradiation conformationnelle et 520 par irradiation interstitielle [5]. Le taux de survie globale de ces patients était de 96 % à 5 ans (IC = [94-98]) et de 82 % à 8 ans (IC = [68-90]). Une rechute biologique définie selon les critères de l'*American Society for Therapeutic Radiology and Oncology* (nadir de PSA + 2 ng/mL), locale ou métastatique, a été détectée chez 14 % des patients à 5 ans (IC = [11-17]) et chez 28 % des patients à 8 ans (IC = [21-36]).

Modèles d'analyse

Analyse trop souvent classique

Pour mettre en évidence l'impact péjoratif de la rechute sur le pronostic vital des patients, on peut être tenté de regrouper les patients en deux groupes, les patients avec rechute (N = 85) et les patients sans rechute (N = 666) à la date de dernières nouvelles. L'estimation des taux de survie à 5 ans par la méthode de Kaplan-Meier suggère l'absence de différence sur la survie globale des patients à 5 ans entre les patients ayant rechuté (96 %, IC = 94-98) et les patients sans rechute à la date de dernières nouvelles (97 %, IC = 89-99) *(figure 1)*.

Pour comparer deux groupes dans une analyse de survie, on est censé connaître l'appartenance à l'un des deux groupes au début de l'analyse. Ce n'est pas le cas d'une rechute qui peut apparaître après un certain temps. Comme le statut de rechute n'est pas connu au début, une analyse classique souffre d'un biais de sélection important en classant les patients décédés très précocement dans le seul groupe des patients sans rechute [6]. En réalité, le statut de la rechute pour ces patients de très mauvais pronostic est manquant à la date de dernières nouvelles car le suivi insuffisant et l'état de santé dégradé de ces patients ne permettent pas l'obtention des examens nécessaires au diagnostic d'une rechute. Ces arguments sont transposables à toute covariable pour laquelle la valeur n'est pas connue à la date d'origine de la survie étudiée, comme la réponse au traitement, par exemple. Malheureusement, il y a encore beaucoup de publications qui analysent la réponse au traitement comme un facteur pronostique de la survie par ces méthodes incorrectes [7].

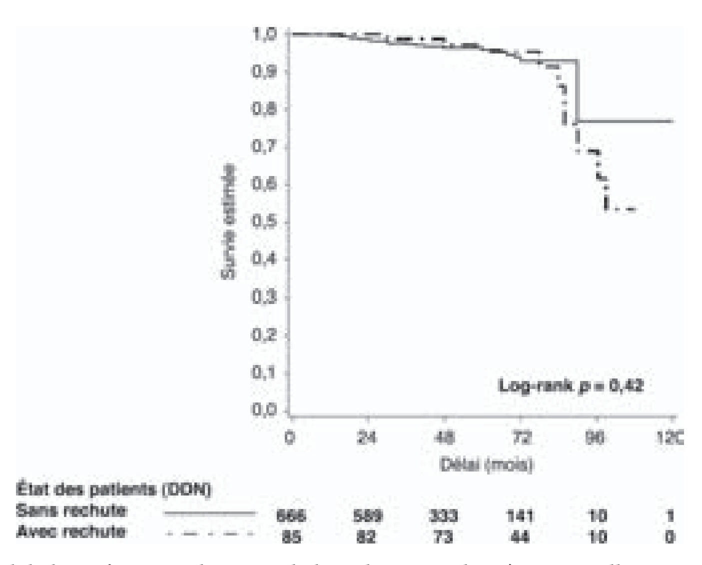


Figure 1. Survie globale en fonction du statut de la rechute aux dernières nouvelles.
DDN : date des dernières nouvelles.

Analyse par la méthode « landmark »

Pour limiter ce biais, on a recours le plus souvent à une analyse par « landmark » qui compare les patients en fonction du statut de l'événement intermédiaire à un temps de référence, le temps de landmark. Les décès des patients observés ou censurés avant le temps de landmark sont exclus de l'analyse. Par exemple sur les 671 patients de notre cohorte encore à l'étude (ni décédés, ni censurés) 24 mois après l'initiation du traitement (M24), on observe une différence de 15 % à 5 ans (98 % *vs* 83 %) et de 49 % à 8 ans (86 % *vs* 35 %) en faveur des patients vivants et sans rechute à M24 *(figure 2)*. La statistique de test du log-rank met en évidence un impact péjoratif du statut de la rechute à M24 sur le pronostic vital des patients ($p < 0{,}0001$). Contrairement à l'analyse précédente, cette analyse a permis de s'affranchir des biais de sélection et confirme le résultat attendu : la détérioration du pronostic vital des patients après une rechute qui a été observée dans les premiers 24 mois. Les résultats de l'analyse par landmark restent cependant conditionnés au choix d'un temps de référence souvent arbitraire et classent les patients qui rechutent après le temps de landmark choisi dans la catégorie de patients sans rechute.

Analyse par la méthode de Mantel-Byar

Une méthode non paramétrique pour comparer la survie de deux groupes où l'appartenance à l'un ou l'autre groupe dépend du temps ($Z(t) = 0,1$) a été proposée par Mantel et Byar [8]. On s'intéresse, par exemple, à tester si l'apparition d'une rechute modifie le risque de décès. Pour un patient qui ne rechute jamais, le vecteur Z(t) aura « 0 » comme valeur pendant toute la durée de

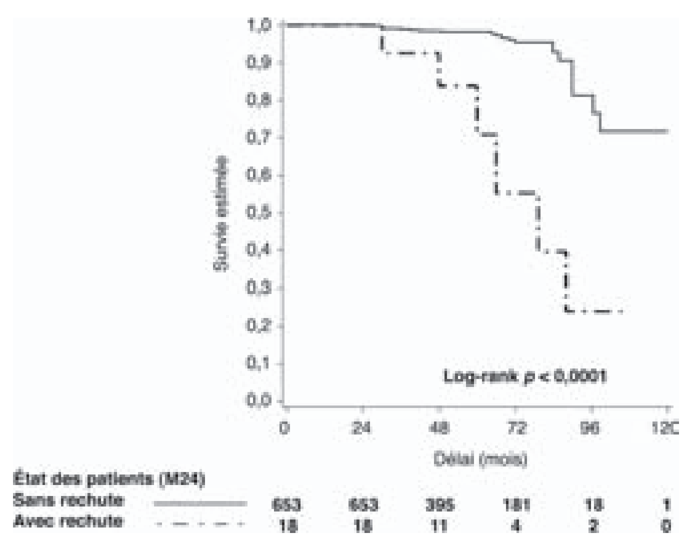


Figure 2. Survie globale : analyse par Landmark 24 mois après l'initiation du traitement.
M24 : 24 mois après l'initiation du traitement.

son suivi. Pour un patient qui rechute au temps t, le vecteur Z(t) aura « 0 » comme valeur pour tous les délais de 0 à t^- (c'est-à-dire, juste avant la rechute) et le vecteur Z(t) aura « 1 » comme valeur à partir du temps t jusqu'au décès ou dernières nouvelles.

Soit T le délai de survie d'un patient, on distingue par $h_0(t)$ la fonction de risque instantané de décès d'un patient avant rechute $(Z(t) = 0)$

$$h_0(t) = P\left[T \in [t, t+dt) \mid T \geq t, Z(t) = 0\right]$$

et $h_1(t)$ la fonction de risque instantané de décès d'un patient après rechute $(Z(t) = 1)$,

$$h_1(t) = P\left[T \in [t, t+dt) \mid T \geq t, Z(t) = 1\right]$$

Pour tester l'hypothèse nulle $H_0 : h_0(t) = h_1(t)$, Mantel et Byar ont proposé le calcul de la statistique suivante :

$$LR_{MB} = [O - E]^2 / V$$

où O désigne le nombre total des décès observés après une rechute, E le nombre total de décès après rechute attendus sous H_0. Cette statistique suit une loi de chi-2. Soit $t_1 < ... < t_K$ la séquence ordonnée des temps de décès observés, d_j et r_j les nombre de décès observés et patients à risque au temps t_j, le nombre de décès attendus E est ici rapporté au nombre de patients $r_{1,j}$ encore à l'étude et ayant rechuté dans l'intervalle $(0,t_j)$,

$$E = \sum_{j=1}^{K} r_{1,j} \frac{d_j}{r_j}.$$

La variance de la différence O - E est estimée par V,

$$V = \sum_j V_j = \sum_j d_j r_{1,j} (r_j - r_{1,j}) r_j^{-2} (r_j - d_j)(r_j - 1)^{-1}.$$

La statistique de Mantel-Byar appliquée aux données de notre exemple est égale à 6,382 et confirme l'impact attendu de la rechute sur la survie globale ($p = 0{,}0115$).

Analyse par le modèle de Cox avec une covariable dépendant du temps

Le modèle de Cox est un modèle de régression semi-paramétrique [1] pour des données de survie censurées à droite qui suppose la proportionnalité des fonctions de risque instantané. La valeur d'une covariable étudiée peut changer au cours du suivi. Le modèle de Cox peut gérer cette situation et il devient :

$$h(t|Z_i(t)) = h_0(t)e^{\beta' Z_i(t)} \quad (1)$$

où $h_0(t)$ désigne une fonction de risque instantané arbitraire, β un vecteur de coefficients de régression et $Z(t)$ un vecteur de variables indépendantes mesurées au cours du temps,

$$\beta = (\beta_1, \ldots, \beta_p)', Z_i(t) = (Z_{i1}(t), \cdots, Z_{ip}(t))'$$

Soit R_j l'ensemble des labels des individus ni décédés ni censurés dans l'intervalle $(0,t_j]$, D_j l'ensemble des labels des individus décédés au temps t_j, le vecteur des paramètres est estimé par maximum de vraisemblance partielle,

$$\beta^* = \operatorname{argmax} L(\beta)$$

où $L(\beta)$ désigne la fonction de vraisemblance partielle [9]

$$L(\beta) = \prod_{j=1}^{K} \frac{e^{\beta' \sum_{l \in R_j} e^{\beta' Z_l(t_j)}}}{\left[\sum_{l \in R_j} e^{\beta' Z_l(t_j)}\right]^{d_j}}$$

Soit $U(\beta) = \partial logL(\beta)/\partial\beta$ la statistique du score, l'estimateur de la variance de β^* est obtenue en prenant l'inverse de la matrice d'information de Fisher $I(\beta) = \partial U(\beta)/\partial\beta$ évaluée en $\beta = \beta^*$,

$$V(\beta) = I^{-1}(\beta^*).$$

Sous l'hypothèse nulle $H_0 : \beta = 0$, la statistique de test du score

$$S = U(0)' I^{-1}(0) U(0)$$

suit une loi de χ^2 avec p degrés de liberté. Si l'on suppose que les variables $Z(t)$ prennent leurs valeurs dans l'ensemble {0,1}, on a les relations suivantes :

$$U(0) = \sum_j U_i = \sum_j d_{1,j} - d_j r_j^{-1}$$

et

$$I(0) = \sum_j d_j r_{1,j}(r_j - r_{1,j}) r_j^{-2}.$$

En l'absence de décès multiples *($d_j = 1$)*, les tests du score et de Mantel-Byar sont rigoureusement identiques. En pratique, il est possible de substituer à la procédure de Mantel Byar le résultat du test du score dérivé du modèle (1) en posant :

- $Z_i(t) = 1$ pour tous les délais supérieurs au délai de rechute du patient ;
- $Z_i(t) = 0$, sinon.

Appliqué aux données de l'exemple, le seuil de significativité du test du score dérivé du modèle est égal à 0,0117, une valeur proche du résultat de la procédure Mantel Byar ($p = 0{,}0115$). Les éléments nécessaires pour le calcul sont présentés dans le *tableau I*. La contribution aux éléments de calcul de la statistique suit le même principe que la statistique de test du log-rank habituelle (*cf.* chapitre III.1 « Données de survie », page 129). La seule différence consiste à identifier, à chaque décès, le nombre d'individus à risque dans chaque groupe. Ce nombre diminue au cours

Tableau I. Calcul de la statistique de Mantel-Byar.

	Groupe avec rechute $Z(T_j) = 1$			**Groupe sans rechute $Z(T_j) = 0$**				
T_j	$r_{1,j}$	$d_{1,j}$	$e_{1,j}$	$r_{0,j}$	$d_{0,j}$	$e_{0,j}$	I_j	V_j
0,70	0	0	0,000	751	1	1,000	0,000	0,000
3,9	0	0	0,000	737	1	1,000	0,000	0,000
14,1	10	0	0,014	706	1	0,986	0,014	0,014
16,3	10	0	0,014	693	1	0,986	0,014	0,014
18,0	12	0	0,017	678	1	0,983	0,017	0,017
19,0	13	0	0,019	676	1	0,981	0,019	0,019
20,0	14	0	0,020	674	1	0,980	0,020	0,020
21,4	14	0	0,021	659	1	0,979	0,020	0,020
...	...	...	...	...	...	...	...	...
85,8	15	1	0,357	27	0	0,643	0,230	0,230
88,0	15	2	0,732	26	0	1,268	0,464	0,452
90,1	9	0	0,429	12	1	0,571	0,245	0,245
96,6	9	1	0,450	11	0	0,550	0,248	0,248
97,0	8	0	0,421	11	1	0,579	0,244	0,244
97,1	8	1	0,444	10	0	0,556	0,247	0,247
101,0	7	1	0,438	9	0	0,563	0,246	0,246
TOTAL		10	5,220		25	29,780	3,591	3,580

du temps dans le groupe sans rechute et il augmente dans le groupe avec rechute au fur et à mesure que les rechutes sont observées. Par exemple, au temps t = 14,1, 10 patients ont eu une rechute entre les temps 3,9 et 14,1 mois et sont donc à risque de décès dans ce groupe à partir de ce moment-là. Au temps t = 14,1, 706 patients sont à risque de décès dans le groupe sans rechute car vivants sans rechute et non perdus de vue.

Pour cet exemple, les deux statistiques sont assez poches avec :

$S = (10 - 5{,}220)^2/3{,}591 = 6{,}363$;

et

$LR_{MB} = (10 - 5{,}220)^2/3{,}580 = 6{,}382$.

Le modèle de régression (1) généralise la procédure de Mantel-Byar dans le cas d'une variable continue *Z(t)* ou d'un vecteur *Z(t)* dont certaines composantes peuvent dépendre du temps. Il permet d'estimer l'effet d'une covariable dépendant du temps et d'ajuster le résultat du test de Mantel-Byar sur le niveau d'une ou de plusieurs variables pronostiques de la survie étudiée. En revanche, le modèle ne permet pas de prédire la survie future de patients si le vecteur de ces covariables dépend du temps [9].

Prédiction dynamique par analyse « landmark » à un temps donné

En supposant un faible pourcentage de patients perdus de vue à un horizon donné t_{hor}, une approximation raisonnable de la survie des patients au temps t_{hor} sachant $T > t_L$ [11, 12]

$$S(t_{hor} | Z_L, T > t_L) = P(T > t_{hor} | Z_L, T > t_L)$$

est obtenue en supposant un modèle avec risques proportionnels et une analyse « landmark »

$$h_L(t | Z_L) = h_{0L}(t) e^{\beta_L Z_L}, \; t_L < t \leq t_{hor} \quad (2)$$

où seuls les patients encore à l'étude au temps t_L et les décès survenus dans l'intervalle $[t_L, t_{hor}]$ contribuent à l'estimation du modèle (les autres décès sont censurés par l'analyse au temps t_{hor}). La valeur prédite par le modèle (2) est obtenue en substituant aux paramètres du modèle β_L et $h_{0L}(t)$ les estimateurs classiques [10] :

$$\hat{S}(t_{hor} | Z_L, T > t_L) = \exp\left(- \int_{t_L}^{t_{hor}} d\hat{H}_{0L}(t) \right)^{\exp(\hat{\beta}_L Z_L)}$$

En faisant varier le temps de landmark, la représentation graphique de la valeur prédite

$$\hat{S}(t_{hor} | Z_L, T > t_L)$$

par les différents modèles (2) permet de mieux appréhender l'évolution de la survie à un horizon donné t_{hor} en fonction du délai d'apparition de la rechute par exemple.

Dans le cadre du suivi de notre cohorte, les probabilités estimées de survie 5 ans après l'initiation de la radiothérapie sachant T > t_L, t_L = *M6, M9, ..., M36 (figure 3)* suggèrent un pronostic à 5 ans identique pour les patients sans rechute dans les 3 ans. En revanche, pour les patients qui présentent une rechute dans les 3 ans, le pronostic à 5 ans est d'autant plus défavorable que le délai de rechute est précoce (de 30 % pour une rechute dans les 6 premiers mois à près de 95 % pour une rechute dans les 36 premiers mois). Cette analyse pose cependant le problème du choix du modèle et donc du choix du temps de landmark. van Houwelingen propose une modélisation dynamique pour différentes valeurs $t_L = t_{L1}, ..., t_{Lm}$ [11]

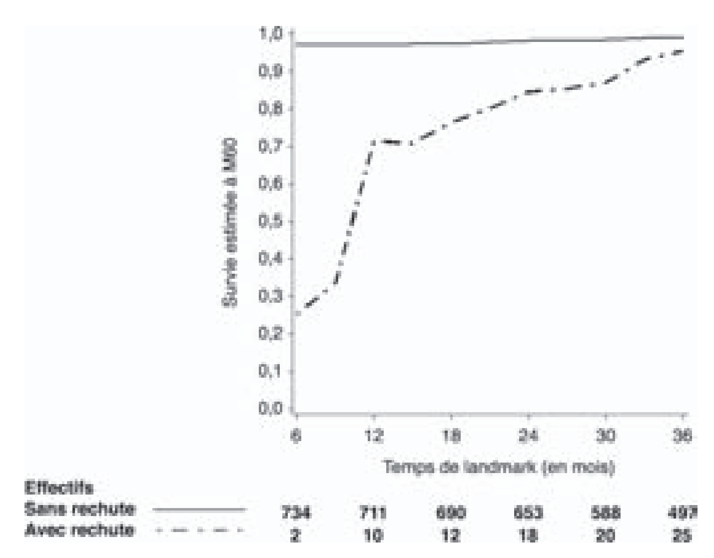


Figure 3. Survie prédite pour différents temps de landmark (t_{hor} = M60).
M60 : 60 mois après l'initiation de la radiothérapie.

$$h_L(t|Z_L) = h_0(t)\exp\left(\gamma_0 s + \gamma_1 s^2 + \beta_0 Z_L + \beta_1 s \times Z_L\right) \quad t_L < t \leq t_{hor}, s = t - t_L \qquad (3)$$

qui permet de s'affranchir d'un choix unique d'un temps de landmark. Les valeurs prédites par le modèle en substituant aux paramètres du modèle (γ_{01}, γ_1,β_0,β_1) et $h_0(t)$ les estimateurs classiques d'un modèle de Cox sont représentées sur la *figure 4*.

Le modèle (3) met en évidence un effet du statut de la rechute sur la survie 5 ans après l'initiation de la radiothérapie avec un risque accru pour les patients présentant une rechute au temps de landmark t_L = *M6, M9, ..., M36* qui décroît au cours du temps (de HR = 25,5 pour un patient avec rechute dans les 6 premiers mois à HR = 4,5 pour un patient avec rechute dans les 36 premiers

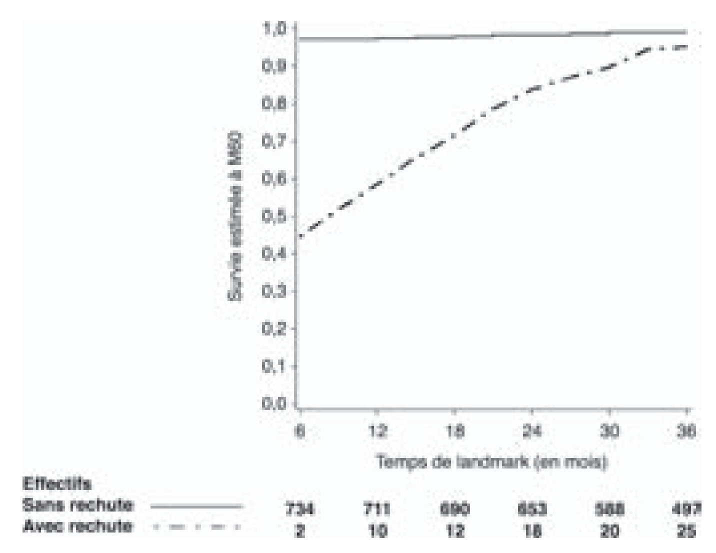


Figure 4. Survie prédite par un modèle « landmark » dynamique (t_{hor} = M60).
M60 : 60 mois après l'initiation de la radiothérapie.

mois). En règle générale, les prédictions du modèle sont biaisées si la proportion de patients perdus de vue à la date t_{hor} fixée est importante. Il est possible de corriger ce biais en intégrant dans le modèle des poids inversement proportionnels à la censure [11, 13].

Ajustement des différents modèles

Ce paragraphe détaille les instructions nécessaires pour réaliser l'ajustement des modèles de régression (1), (2) et (3) discutés dans ce chapitre. Le fichier étudié (PROSTATE) inclut l'enregistrement des variables individuelles suivantes : l'identifiant du patient (SUBJID), le délai avant rechute si renseigné (RELD), le délai avant décès si renseigné (DCD), le délai avant censure ou décès (TEMPS), l'indicateur de censure ou de décès (STATUT = 1 si décès, = 0 sinon). Les analyses ont été réalisées avec le logiciel SAS version 9.2 [14].

Ajustement du modèle (1)

Les instructions suivantes permettent d'ajuster le modèle (1) avec une variable dépendant du temps, Z(t) = 1 si le patient a eu une rechute au temps t, 0 sinon.

```
PROC PHREG DATA = PROSTATE ;
MODEL TEMPS*STATUT(0)= Zt ;
Zt = 0 ; If (RELD ^=. And RELD <= TEMPS) Then Zt = 1 ;
RUN ;
```

Comme attendu, les valeurs de p des statistiques du score ($p = 0{,}0117$) et de Wald ($p = 0{,}0143$) dérivées de ce modèle sont semblables à la statistique introduite par Mantel-Byar ($p = 0{,}0115$).

Ajustement du modèle (2)

Les instructions suivantes permettent de tester l'influence du statut de la rechute connu à M24 noté Z_L en appliquant la méthode du landmark. Le délai de 24 mois après l'initiation de la radiothérapie est la date de landmark choisie (*cf*. section « Analyse par la méthode landmark » supra).

```
PROC PHREG DATA = PROSTATE ;
WHERE TEMPS >= 24 ;
MODEL TEMPS*STATUT(0)= ZL ;
ZL = 0 ; If (RELD ^=. And RELD <= 24) Then ZL = 1 ;
RUN ;
```

La clause WHERE permet d'exclure de l'analyse les décès précoces ou les patients ayant un suivi insuffisant (TEMPS ≤ 24). Les valeurs de p des statistiques du score ($p < 0,0001$) et de Wald ($p < 0,0001$) dérivées de ce modèle confirment le résultat obtenu par la statistique de test du log-rank comparant les deux groupes de patients, $Z_L = 0$ et $Z_L = 1$.

Ajustement du modèle (3)

Pour ajuster le modèle (3), il est nécessaire de regrouper dans un fichier unique les données des analyses « landmark » aux différents t_L temps choisis en censurant les patients à la date t_{hor} choisie, temps > 60.

```
DATA LMS ; SET PROSTATE ;
IF TEMPS > 60 THEN DO ; TEMPS = 60 ; STATUT = 0 ; END ;
DO TL = 6 TO 36 BY 3 ;
IF TEMPS > TL THEN DO ; ZL = 0 ; IF (RELD^=. And RELD <= TL) THEN ZL = 1 ; OUTPUT ; END ;
END ;
RUN ;
```

L'ajustement du modèle est possible dès lors que la procédure utilisée autorise les entrées tardives (option ENTRY) et un estimateur robuste de la variance (option COVS) prenant en compte la corrélation des données attachées à un même individu dans le calcul des statistiques de tests.

```
PROC PHREG DATA = LMS COVS(AGGREGATE) ;
MODEL TEMPS*STATUT(0)= S S2 ZL ZLS / ENTRY = TL ;
S = TL-6 ; S2 = (TL-6)**2 ; ZLS = ZL*S ;
ID SUBJID ;
RUN ;
```

Conclusions

L'analyse trop souvent « classique » qui consiste à comparer la survie des patients en fonction du statut d'un événement survenu après le début de l'étude est biaisée. Dans l'exemple traité, cette analyse incorrecte aboutit à la conclusion inattendue de l'absence de modification du risque de décès après

une rechute. L'analyse par landmark permet de limiter le biais de sélection et de montrer que les rechutes précoces ont un impact plus péjoratif sur la survie que les rechutes tardives et ce, de manière significative. La procédure de Mantel-Byar conduit à la même conclusion en évitant le choix arbitraire d'un temps de démarrage « landmark ». Le modèle de régression de Cox est un outil flexible permettant de généraliser le test de Mantel-Byar au cas d'une variable continue et non plus catégorielle. Il permet également de gérer des situations diverses et variées : évaluer l'impact pronostique d'un marqueur mesuré en cours d'étude indépendant de la valeur du marqueur mesuré au diagnostic ou avant traitement. La prédiction dynamique par landmark proposée par van Houwelingen est un compromis entre ces deux approches, la méthode landmark et la modélisation par un modèle de Cox. Il permet notamment de prédire l'évolution de la survie à un horizon donné t_{hor} en fonction du statut d'un patient Z(t) au cours du temps sans faire d'hypothèses sur l'évolution de Z(t).

Références

1. Cox DR. Regression models and life tables (with discussion). *J R Stat Soc Ser B* 1972 ; 34 : 187-220.

2. Aalen OO. Nonparametric inference for a family of counting processes. *Ann Stat* 1978 ; 6 : 534-45.

3. Andersen PK, Gill RD. Cox's regression model for Counting processes: A large sample study. *Ann Stat* 1982 ; 10 : 1100-20.

4. Therneau TM, Grambsch TM. *Modeling survival Data: Extending the Cox model.* New-York : Springer-Verlag, 2000, 350 pages.

5. Farnault B, Duberge T, Salem N, *et al.* La curiethérapie de prostate par implants permanents peut-elle représenter une alternative à la radiothérapie externe pour les cancers de prostate localisés de risque intermédiaire ? *Cancer/Radiothérapie* 2009 ; 13 : 686.

6. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumour response. *J Clin Oncol* 1983 ; 1 : 710-719.

7. Hill C. Faut-il comparer la survie des répondeurs à celle des non-répondeurs ? *Bull Cancer* 1993 ; 80 : 294-8.

8. Mantel N, Byar DP. Evaluation of response-time data involving transient states: An illustration using heart-transplant data. *J Am Stat Assoc* 1974 ; 69 ; 81-6.

9. Fisher LD, Lin DY. Time-dependent covariates in the Cox proportional-hazards regression model. *Annu Rev Public Health* 1999 ; 20 : 145-57.

10. Breslow NE. Discussion following "Regressions models and life tables" br DR Cox. *J R Stat Soc Ser B* 1974 ; 34 : 187-220.

11. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics* 2007 ; 34 : 70-85.

12. van Houwelingen HC, Putter H. Dynamic predicting by landmarking as an alternative for multi-state modeling: An application to acute lymphoid leukemia data. *Lifetime Data Analysis* 2008 ; 14 : 447-63.

13. Xu R, O'Quigley J. Estimating average regression effect under non-proportional hazards. *Biostatistics* 2000 ; 1 : 423-39.

14. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

Courbes de survie ajustées par des covariables

A. Kramar, M. Velten

Dans les essais thérapeutiques qui comparent le délai jusqu'à la survenue d'un événement tel que le décès, il est souvent nécessaire d'ajuster la comparaison pour tenir compte d'un ou plusieurs facteurs pronostiques importants dans la comparaison des traitements. Si le critère de jugement est un critère de survie, l'analyse la plus adaptée consiste à utiliser soit le test du log-rank ajusté s'il y a peu de facteurs d'ajustement (ou de stratification), soit un modèle plus général tel que le modèle de Cox (*cf.* chapitre IV.2 « Modèle de Cox et index pronostique », page 213). Les résultats de tels essais présentent fréquemment les courbes de survie, pour chaque groupe de traitement, de façon brute et non « corrigée », c'est-à-dire estimées sans prendre en compte une éventuelle différence de répartition des covariables dans les différents groupes de traitement, alors que les tests statistiques sont souvent ajustés [1-3].

L'exemple présenté ci-dessous montre que la comparaison des délais de survie entre deux groupes n'est pas significative, alors que le test stratifié sur un facteur pronostique important permet de mettre en évidence une différence statistiquement significative. Présenter des courbes non ajustées peut donner une fausse impression de la différence entre les groupes. Dans cette situation, il vaut mieux présenter les courbes de survie ajustées.

Méthodes

Analyse classique

Dans le cadre de l'analyse d'un essai thérapeutique randomisé dont le critère de jugement est le délai de survie depuis la date de randomisation, on dispose d'un échantillon de N sujets répartis de manière aléatoire en G groupes de traitements de taille à peu près égale (le plus souvent G = 2).

On caractérise la distribution des temps de survie de chaque groupe de traitement g (g = 1, 2, ..., G) par $S_g(t)$, la probabilité de survie au-delà du temps t. Cette fonction de survie est généralement estimée par la méthode de Kaplan-Meier (*cf.* chapitre III.1 « Données de survie », page 129) [4].

$$\hat{S}_g(t) = \prod_{t_i \leq t} \left[1 - \frac{d_g(t_i)}{n_g(t_i)} \right]$$

où t_i représente les instants ordonnés où des décès sont observés dans le groupe g ; $d_g(t_i)$ et $n_g(t_i)$ représentent respectivement pour le groupe g, le nombre de décès observés en t_i et le nombre de sujets en vie juste avant t_i (donc exposés au risque de décès en t_i).

Analyse ajustée

Pour prendre en compte les facteurs pronostiques dans ces estimations, on procède de la manière suivante : soit Z le vecteur des covariables retenues, on calcule l'estimateur de Kaplan-Meier de $S_g^C(t,z)$ dans chaque groupe de traitement g et pour chaque combinaison z des valeurs des variables pronostiques.

L'estimateur de la survie ajustée est alors donné par la formule suivante :

$$S_g^C(t) = \sum_z \hat{S}_g(t,z)w(z)$$

où $w(z)$ correspond à la fréquence des sujets de profils z dans une population de référence. L'estimation de la variance de l'estimateur de la survie ajustée est donnée par :

$$\mathrm{var}(S_g^C(t)) = \sum_z w^2(z) Var\left[\hat{S}_g(t,z)\right]$$

L'estimateur de la survie ajustée prend en compte une répartition éventuellement différente des covariables dans les groupes de traitements randomisés. Ce calcul correspond à une standardisation directe des taux, par analogie avec les méthodes de standardisation développées en épidémiologie et en santé publique pour permettre des comparaisons valides entre différents groupes qui diffèrent quant aux caractéristiques liées à la survie, l'âge et le sexe la plupart du temps. Ces techniques de standardisation sont toujours utilisées quand il s'agit de comparer la survie à 5 ans entre des populations de plusieurs pays, par exemple, car il faut tenir compte de la répartition par sexe et par âge qui sont des facteurs liés à la survie. Pour l'analyse des essais en cancérologie, on présente les résultats en termes de *hazard ratio* univarié entre les groupes, mais aussi en multivarié ajusté sur des facteurs pronostiques importants. Bien que des techniques d'ajustement soient utilisées pour mesurer l'effet du traitement à niveau égal des facteurs pronostiques, on constate en pratique que les courbes de survie sont le plus souvent présentées sans ajustement.

Remarque : On peut choisir pour population de référence, soit la population de l'essai, soit une autre population de référence et prendre, par exemple, une répartition uniforme des covariables. Enfin, si l'un des groupes de traitement est un groupe contrôle, il peut être intéressant de standardiser tous les autres groupes traités sur celui-ci.

Exemple

L'exemple présenté ici est extrait de l'annexe 3 de l'article de Peto *et al.* [5]. Le *tableau I* présente les délais de survie pour les deux groupes de traitement selon le niveau de la fonction rénale.

Dans cet exemple, la comparaison entre les deux groupes de traitement A et B produit un résultat non significatif (*figure 1*, test du log-rank, $p = 0{,}25$) (*cf.* chapitre III.1, page 129). Le *hazard ratio* est égal à 0,56 (IC95 % : 0,21-1,53, modèle de Cox) (*cf.* chapitre IV.2, page 213).

En revanche, la fonction rénale est un facteur pronostique important. La comparaison des distributions de survie des sujets à fonction rénale normale ou pathologique donne un résultat très significatif (*figure 2*, $p < 0{,}001$). Dans cet exemple, on peut remarquer que dans le groupe A, les patients ayant une fonction rénale pathologique meurent tous avant les patients ayant une fonction rénale normale.

Tableau I. Exemple fil rouge.

Fonction rénale	Groupe A (n = 13)	Groupe B (n = 12)
Normale (n = 18)	70, 76, 180, 195, 210, 632, 700, 1296, 1 990*, 2 240*	8, 220, 365*, 852*, 1 296*, 1 328*, 1 460*, 1 976*
Pathologique (n = 7)	13, 18, 23	8, 52, 63, 63

* Données censurées.

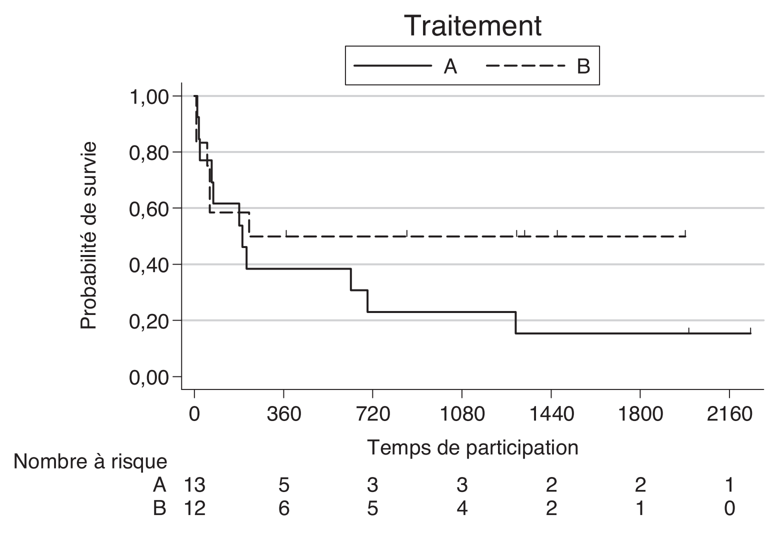


Figure 1. Survie par groupe de traitement (d'après [5]).

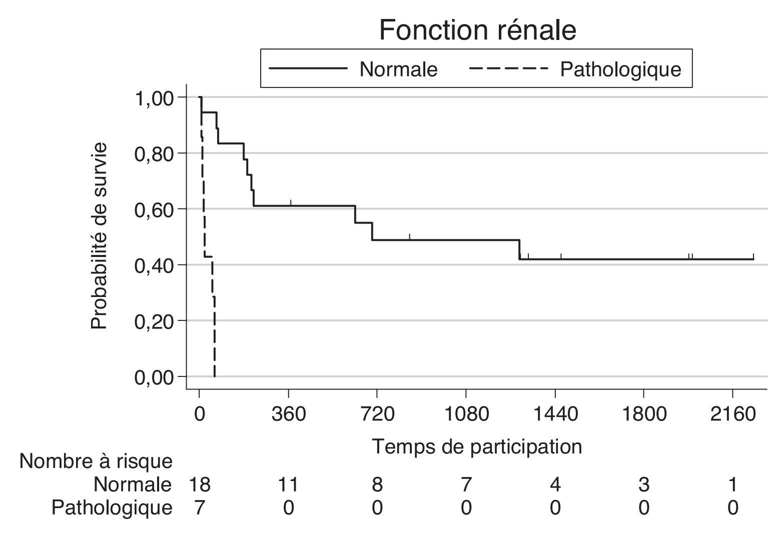


Figure 2. Survie et fonction rénale (d'après [5]).

La comparaison des distributions de survie observées dans les deux groupes de traitement A et B doit donc prendre en compte ce facteur pronostique important. Le résultat du test du log-rank stratifié sur la fonction rénale est significatif ($p = 0{,}016$). Le *hazard ratio* est égal à 0,23 (IC95 % : 0,06-0,84, modèle de Cox) (*cf*. chapitre IV.2, page 213). On peut donc conclure que les distributions de survie des deux groupes de traitement sont significativement différentes à fonction rénale constante.

On est confronté ici à la situation où on fait apparaître une différence significative en tenant compte d'un facteur qui est la fonction rénale. Pour faire correspondre les estimations des taux de survie avec les résultats des tests statistiques, il est important de présenter les estimations des taux de survie ajustés.

Les taux de survie observés pour chaque traitement et pour chaque type de fonction rénale, sont présentés dans le *tableau II* et estimés d'après la méthode de Kaplan-Meier. Pour obtenir l'estimation de la survie ajustée, $S_g^C(t)$ à un instant t donné, on calcule la moyenne des survies $S_g(t,z)$ observées à l'instant t dans ce groupe de traitement, pour chaque niveau de la fonction rénale, pondérée par la distribution $w(z)$ prise comme référence (d'après [1]).

Si l'on prend pour référence la population de malades définie par le protocole de l'essai, on estime $w(z)$ sur l'ensemble des patients inclus : 13 sujets dans le groupe A dont 3 avec une fonction rénale pathologique (P) et 12 sujets dans le groupe B dont 4 avec une fonction rénale pathologique. On prendra pour $w(z = P)$ la valeur 7/25 = 0,28, et pour $w(z = N)$ la valeur complémentaire 0,72.

À titre d'exemple, on calcule, à partir du *tableau II*, l'estimation de la survie ajustée dans le groupe B *à* 52 jours :

$$S_B^c(52) = \hat{S}_B(52,P) * w(z = P) + \hat{S}_B(52,N) * w(z = N)$$

= (0,50) * (0,28) + (0,875) * (0,72) = 0,77

et la variance dans le groupe B à 52 jours :

$$Var\left[S_B^c(52)\right] = w^2(z = P) * Var\left[\hat{S}_B(52,P)\right] + w^2(z = N) * Var\left[\hat{S}_B(52,N)\right]$$

= 0,0784 * 0,0625 + 0,5184 * 0,0137 = 0,012

Cette valeur correspond à un écart-type qui vaut 0,109. On obtiendrait de même la survie ajustée du groupe B aux autres instants de décès 8, 63, 220. Les estimations des taux de survie observés dans chaque groupe pour chaque groupe pronostique N et P, ainsi que les taux de survie ajustés corrigés sont présentées dans le *tableau II*.

Tableau II. Estimation des taux de survie par groupe et fonction rénale.

	A				B			
Temps	$S_A(t)$	**N**	**P**	$S_A^c(t)$	$S_B(t)$	**N**	**P**	$S_B^c(t)$
8	1,000	1,0000	1,0000	1,000	0,833	0,8750	0,7500	0,770
13	0,923	1,0000	0,6667	0,907	0,833	0,8750	0,7500	0,770
18	0,846	1,0000	0,3333	0,813	0,833	0,8750	0,7500	0,770
23	0,769	1,0000	0,0000	0,720	0,833	0,8750	0,7500	0,770
52	0,769	1,0000	–	0,720	0,750	0,8750	0,5000	0,770
63	0,769	1,0000	–	0,720	0,583	0,8750	0,0000	0,630
70	0,692	0,9000	–	0,648	0,583	0,8750	–	0,630
76	0,615	0,8000	–	0,576	0,583	0,8750	–	0,630
180	0,539	0,7000	–	0,504	0,583	0,8750	–	0,630
195	0,461	0,6000	–	0,432	0,583	0,8750	–	0,630
210	0,385	0,5000	–	0,360	0,583	0,8750	–	0,630
220	0,385	0,5000	–	0,360	0,500	0,7500	–	0,540
632	0,308	0,4000	–	0,288	0,500	0,7500	–	0,540
700	0,231	0,3000	–	0,216	0,500	0,7500	–	0,540
1 296	0,154	0,2000	–	0,144	0,500	0,7500	–	0,540

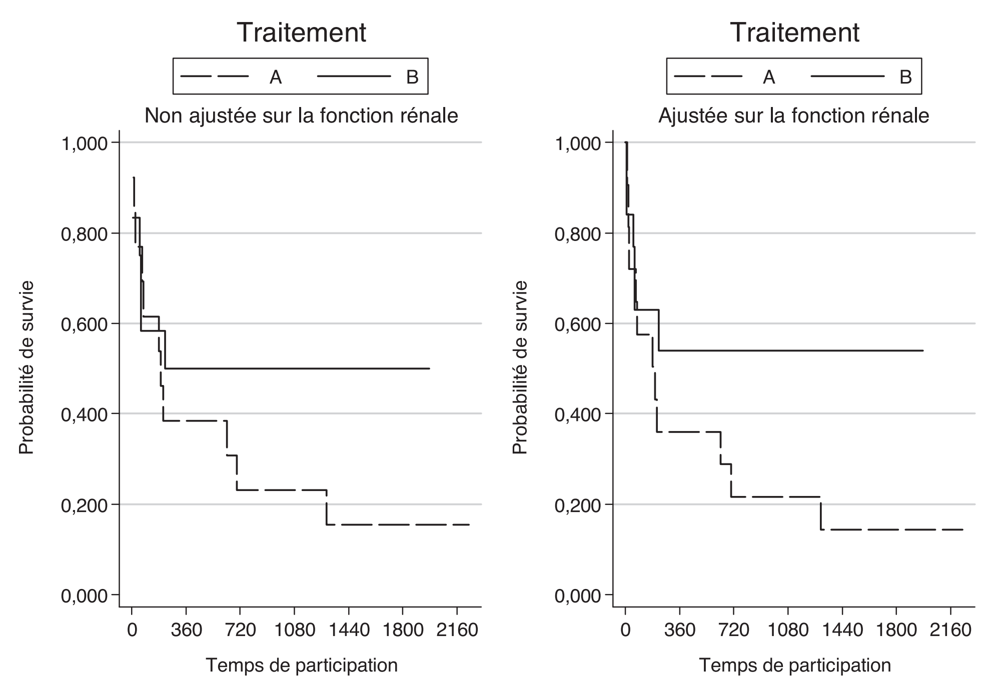


Figure 3. Survie observée et ajustée par groupe de traitement (d'après [2]).

Bien que peu visible à l'œil nu *(figure 3)*, l'écart entre les deux courbes ajustées (courbes de droite) est plus grand que l'écart entre les courbes non ajustées (courbes de gauche), et cet écart est suffisant pour rendre la différence statistiquement significative. Par exemple à 210 jours, la survie diminue et passe de 0,385 à 0,360 dans le groupe A alors qu'elle augmente et passe de 0,583 à 0,630 dans le groupe B *(tableau II)*.

Remarque : En ce qui concerne le calcul de la variance ajustée, elle n'est plus définie lorsque le dernier sujet à risque décède. La plupart des logiciels donnent pour la variance la valeur zéro. Lorsque l'on ajuste deux courbes de survie, la variance ajustée n'est en toute rigueur définie qu'à condition que la survie observée ne soit pas nulle et ce, dans aucune des catégories des sujets définies par le groupe et les facteurs d'ajustement.

Conclusions

Dans cet exemple, on a présenté une technique permettant de fournir des estimations des courbes de survie ajustées qui correspondent aux résultats des tests ajustés. Les résultats permettent de représenter les estimations qui sont en accord avec les tests statistiques. Dans cet exemple, la comparaison univariée entre les deux groupes n'a pas donné de résultat significatif alors que la comparaison multivariée (ajustée sur la fonction rénale) a montré une différence statistiquement significative. Cela s'explique par le fait que les estimations ajustées dans le groupe B sont plus élevées que les estimations brutes et ce, quel que soit le temps, alors que celles du groupe A sont plus petites. En ajustant sur la fonction rénale, les distributions de survie des groupes de traitement A et B s'écartent davantage et reflètent mieux la différence significative observée entre les traitements à fonction rénale constante.

Références

1. Chang I, Gelman R, Pagano M. Corrected group prognostic curves and summary statistics. *J Chronic Dis* 1982 ; 35 : 669-74.

2. Kramar A, Com-Nougué C. Estimation des courbes de survie ajustées. *Rev Epidemiol Sante Publique* 1990 ; 38 : 149-52.

3. Ghali WA, Quan H, Brant R, *et al.* ; APPROACH (Alberta Provincial Project for Outcome assessment in Coronoary Heart Disease Investigators). Comparison of 2 methods for calculating adjusted survival curves from proportional hazards models. *JAMA* 2001 ; 286 : 1494-7.

4. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *JASA* 1958 ; 53 : 457-81.

5. Peto R, Pike MC, Armitage P. Organisation et analyse des essais thérapeutiques comparatifs comportant une longue surveillance des malades. *Rev Epidemiol Sante Publique* 1979 ; 27 : 167-255.

Méta-analyse d'essais randomisés

S. Chabaud, J.P. Pignon, A. Aupérin

Ce chapitre se veut didactique et donne les éléments minimaux et récents sur les méta-analyses pour comprendre un article. Nous renvoyons le lecteur qui souhaite approfondir cette thématique vers d'autres livres ou références sur ce sujet très développé depuis les années 1990 en cancérologie [1-4].

Une méta-analyse est une synthèse non biaisée des résultats de l'ensemble des essais randomisés étudiant une question similaire. Cette synthèse est quantitative et permet souvent d'aboutir à des conclusions alors que les résultats des essais qui sont inclus dans cette méta-analyse peuvent ne pas être concluants à eux seuls par manque de puissance. Il faut la distinguer d'une simple synthèse de résultats d'essais *(pooled analysis)* qui ne recherche pas l'exhaustivité ou d'une revue systématique *(systematic review)* qui correspond à une recherche exhaustive associée à un bilan systématique de la qualité des essais randomisés répondant à une question donnée mais sans réalisation d'une synthèse quantitative. Cette revue systématique est une étape préalable indispensable à une méta-analyse, implicite à ce type d'études. En l'absence de données suffisantes en quantité et en qualité, une revue systématique peut ne pas déboucher sur une méta-analyse.

L'objectif d'une méta-analyse peut être :
- d'augmenter la puissance du test statistique en augmentant la taille de l'échantillon ou d'améliorer la précision de l'estimation de l'effet du traitement ;
- de lever le doute en cas de résultats apparemment discordants entre essais ou entre revues de la littérature et/ou d'expliquer la variabilité des résultats ;
- d'établir des hypothèses pour un nouvel essai ;
- du fait de la grande quantité d'information accumulée, de réaliser des analyses exploratoires plus approfondies pour identifier des facteurs « patients » ou « essais » pouvant modifier l'effet du traitement.

Il existe deux grands types de méta-analyses, celles basées sur des données individuelles et celles basées sur des données résumées généralement extraites des publications. La méta-analyse sur données individuelles reste la méthode de référence, permettant une vérification et une analyse approfondie des données, mais elle est plus consommatrice de ressources que la méta-analyse de données publiées. La différence de ressources nécessaires entre ces deux types de méta-analyses est fonction principalement du nombre d'essais impliqués, mais aussi de leur ancienneté, plus de moyens pouvant être nécessaires pour recueillir les données de « vieux » essais. La méta-analyse sur données individuelles permet de standardiser le format des données, de faciliter la comparaison des résultats d'un essai à l'autre et de demander la mise à jour du suivi si nécessaire.

Méthodes

Écriture du protocole

Il est indispensable d'écrire un protocole avant de commencer la méta-analyse. Cette étape doit permettre de bien identifier la question posée, le but de l'étude, de définir les critères d'inclusion et la méthode de recherche des essais, les méthodes de recueil des données et de vérification de leur qualité, la stratégie d'analyse et les méthodes statistiques. Afin d'adapter la stratégie d'analyse aux données disponibles, une description des essais éligibles et de leurs caractéristiques (taille, population, traitement, etc.) est utile. En revanche, la rédaction du protocole doit être réalisée sans la connaissance des résultats des essais.

Identification des essais à inclure

Cette étape est très importante. En effet, pour ne pas être biaisée, une méta-analyse doit inclure de manière exhaustive tous les essais cliniques étudiant la question posée. Il ne faut pas se contenter d'inclure dans la méta-analyse uniquement les essais publiés en anglais. En effet, les essais dont les résultats sont statistiquement significatifs sont plus souvent publiés que les essais négatifs, ce qui crée un biais de publication. Les essais négatifs sont plus souvent publiés en langue non anglaise (biais de langage) ou dans des petites revues non référencées dans Medline que les essais positifs. Il est donc important de multiplier les sources de recherche en utilisant plusieurs bases bibliographiques électroniques, les actes des congrès et les registres d'essais cliniques (comme www.clinicaltrial.gov). En raison de l'exigence récente, depuis 2005, du comité international des éditeurs de journaux médicaux *(International Committee of Medical Journal Editors)* d'une déclaration des essais dans un registre, préalablement au début des inclusions, pour pouvoir être ensuite publiés, les registres exhaustifs d'essais se développent, ce qui facilitera la réalisation des méta-analyses. La *Cochrane Collaboration* a entrepris la recherche systématique des essais pour l'ensemble des spécialités médicales afin de constituer un registre mondial. L'industrie pharmaceutique et les agences du médicament peuvent être également utiles. À partir des références identifiées, il faut sélectionner les essais éligibles sur la simple lecture soit du titre de l'article (du protocole si l'essai n'est pas publié) soit de son résumé puis, pour ceux qui semblent être éligibles, sur la lecture des publications. Pour éviter un biais dans la sélection des essais, il est recommandé de faire réaliser cette étape indépendamment par deux personnes avec une tierce personne pour résoudre les discordances afin d'éviter le biais de sélection.

Recueil et vérification des données

Une fois tous les essais identifiés :

- si l'on souhaite faire une méta-analyse sur données individuelles, il est nécessaire de contacter les investigateurs ou les groupes de recherche ou bien les promoteurs des différents essais que l'on souhaite inclure dans la méta-analyse. En effet, l'accord des différents intervenants est indispensable avant de recueillir les bases de données de chaque essai. Une structure de base type est

proposée et une actualisation du suivi souvent demandée. L'étape de transcodage (du format de l'essai à celui commun à tous les essais de la méta-analyse) devra faire l'objet d'un contrôle de qualité ;

• en cas de méta-analyse sur données résumées, la première étape est généralement d'extraire les données des publications (articles et résumés) en s'assurant de disposer de la publication la plus récente et en étant attentif au risque d'inclure deux fois un essai en cas de publications multiples. Les investigateurs devront être contactés si les données à recueillir ne sont pas disponibles dans l'article (biais de report) ou si elles sont disponibles mais dans un format non exploitable pour la méta-analyse ou bien si les données ne sont pas publiées. L'aide d'un traducteur pourra être nécessaire pour éviter d'exclure un essai publié dans une langue non maîtrisée par les personnes en charge de la méta-analyse. Enfin, un protocole et une fiche de recueil des données ainsi qu'une double extraction de celles-ci devront être prévus pour limiter les erreurs et les biais associés à cette étape. Pour les données de survie, il est important d'extraire un *hazard ratio* ou des paramètres permettant de l'estimer [5] plutôt qu'un simple nombre d'événements qui ne permettrait d'estimer qu'un odds ratio, paramètre qui peut être biaisé quand il sert à résumer des données de survie [6].

Une étape importante de la méta-analyse est ensuite de vérifier la qualité de chaque essai et, le cas échéant, de décider, sur des critères objectifs et justifiés, de ne pas tenir compte d'un essai mal conduit dans la méta-analyse. Pour les méta-analyses sur données résumées, l'étude de la qualité des essais ne repose généralement que sur leur publication. Pour les méta-analyses sur données individuelles, la qualité des essais peut être également évaluée à partir des données de l'essai [7]. Les trois points à systématiquement étudier quel que soit le type de méta-analyse sont : analyse en intention de traiter ou non ; qualité du tirage au sort ; qualité du suivi des patients. Une méta-analyse sur données individuelles peut permettre de refaire en intention de traiter une analyse qui n'aurait pas été faite ainsi pour la publication de l'essai. Des grilles comme celle de Jadad ont été proposées pour évaluer la qualité de la publication des essais [8]. Il faut se méfier des scores globaux qui peuvent correspondre à des réalités différentes et rapporter les résultats des différentes composantes de ces scores. Les analyses pondérées sur les scores de qualité sont à proscrire à cause de la difficulté d'interprétation de tels résultats et du risque de résultats discordants en fonction de la grille utilisée [9]. Les essais présentant des biais majeurs doivent être exclus. L'impact des essais de qualité incertaine peut être étudié par une analyse de sensibilité dans laquelle les résultats avec et sans ces essais sont comparés. L'aide d'un comité d'experts indépendants statuant à partir de rapports de qualité ne permettant pas d'identifier l'essai peut être précieux dans ce cas pour décider de la meilleure stratégie.

Enfin, chaque essai peut avoir sa propre définition ou manière de recueillir le critère que l'on souhaite étudier dans la méta-analyse. Il n'y a pas de problème lorsque la méta-analyse porte sur la mortalité globale mais cela peut devenir problématique lorsque l'on souhaite étudier un autre critère. Il faut alors bien vérifier que chaque essai utilisait bien la même définition et les mêmes modalités d'obtention. Le recours aux données individuelles permet de garantir une définition identique et offre la possibilité d'analyser des critères de jugement non systématiquement étudiés dans les publications, tels que l'incidence cumulée de certains événements survenant au cours du temps (progression locale, progression à distance, etc.). On voit apparaître dans les

méta-analyses une diversification des critères de jugement étudiés qui ne se réduisent plus à l'étude de la survie globale. Ainsi, la standardisation des échelles de toxicité permet une étude plus fréquente de celle-ci dans le cadre des méta-analyses.

Analyse principale et représentation graphique

La première étape de l'analyse consiste à estimer l'effet global moyen à l'aide d'un test statistique. La méthode utilisée est une analyse stratifiée sur chaque essai, dont le principe est d'évaluer l'effet propre dans chaque essai puis d'en faire une synthèse quantitative pour aboutir à un résultat global (par ex., test du log-rank stratifié par essai).

Une fois l'effet moyen estimé, un test d'hétérogénéité doit être effectué. En effet, le résultat global obtenu ne peut s'interpréter qu'en l'absence d'hétérogénéité entre les différents essais. On parle d'hétérogénéité lorsque la variation des résultats des essais dépasse une simple fluctuation d'échantillonnage. Les tests d'hétérogénéité étant peu puissants, il faut s'assurer, même dans le cas d'un résultat non significatif, qu'aucun des essais ne présente de résultats extrêmes. Dans le cas de résultats extrêmes, une analyse de sensibilité peut permettre de conforter le résultat observé.

Deux grandes catégories de modèles statistiques peuvent être utilisées pour réaliser une méta-analyse : les modèles à effets fixes ou les modèles à effets aléatoires [2, 10, 11].

Dans un modèle à effets fixes, on considère que chaque essai *i* représente une estimation d'un unique « vrai » effet du traitement $\hat{\theta}i = \theta$. *L'estimation de l'effet commun peut alors être obtenue en utilisant la moyenne des estimations de chaque essai pondérée par l'inverse de leur variance* ωi. *Cette pondération est nécessaire du fait que les différentes estimations de* θi *ne sont pas égales en termes de précision ou de variance* $\sigma i2$. L'effet du traitement commun s'écrit alors :

$$\hat{\theta} = \frac{\sum \theta_i \omega_i}{\sum \omega_i}$$

Si le test d'hétérogénéité est significatif, il faut alors utiliser un modèle à effets aléatoires.

Le modèle à effets aléatoires permet au vrai effet du traitement de varier en faisant l'hypothèse que chaque essai représente une estimation d'un réel effet du traitement θi, lui-même étant une variable aléatoire normalement distribuée autour d'un effet global constant de moyenne θ et de variance $\sigma 2$. Ce modèle permet de décomposer la variance totale en une variabilité inter-essai et une variabilité intra-essai.

En pratique, la recherche d'hétérogénéité statistique doit être systématique, et le recours aux modèles à effets aléatoires est variable d'une équipe à l'autre, en particulier entre celles travaillant sur données individuelles et celles travaillant sur données publiées. Les premières réservent essentiellement les modèles à effets aléatoires aux situations où l'hétérogénéité ne peut être expliquée, ce qui peut se comprendre car, dans ce cas, les possibilités d'explorer la ou les sources de l'hétérogénéité sont plus grandes. Les secondes y recourent plus systématiquement soit comme analyse

principale, soit comme analyse de sensibilité, voire seulement en cas d'hétérogénéité significative. En l'absence d'hétérogénéité, les modèles à effets fixes et aléatoires conduisent aux mêmes résultats et le modèle le plus simple est à recommander.

Les résultats d'une méta-analyse sont présentés sous forme graphique où l'effet du traitement dans chaque essai et l'effet global du traitement sont positionnés sur le même graphique sous la forme de carré entouré par son intervalle de confiance et d'un losange pour l'effet global. Une ligne verticale représentant l'absence d'effet est également tracée *(figure 1)*.

Exploration de l'hétérogénéité

En présence d'hétérogénéité, il faut essayer de l'expliquer. En effet, l'exploration de l'hétérogénéité peut permettre d'identifier des types de traitement plus ou moins efficaces ou des groupes de patients bénéficiant plus ou moins du traitement (voir section suivante pour les méthodes d'analyse). La recherche de(s) l'essai(s) pouvant entraîner l'hétérogénéité est possible en inspectant le graphe des hazard ratios et en utilisant des analyses graphiques [12, 13]. Baujat *et al.* [13] ont proposé une méthode graphique qui permet de visualiser facilement les essais les plus hétérogènes et les plus influents de la méta-analyse sur un graphique en deux dimensions. Mais l'exploration de l'hétérogénéité par des analyses de sous-groupes définis *a posteriori* sur la base de la discordance des résultats observés doit être évitée [14]. La planification *a priori* de l'exploration d'une éventuelle hétérogénéité par comparaison de l'effet du traitement, entre des groupes d'essais définis en fonction des traitements qu'ils testent ou entre des sous-groupes de patients définis en fonction des caractéristiques des patients ou des tumeurs, est préférable. Si des sources d'hétérogénéité sont identifiées, cela permet de générer de nouvelles hypothèses à tester dans des études ultérieures. En présence d'hétérogénéité pour laquelle on ne trouve pas d'explication liée au type de patients/tumeurs inclus dans les essais ou aux traitements étudiés, l'hétérogénéité doit être prise en compte pour estimer un effet global du traitement dans des modèles à effets aléatoires où l'effet du traitement réel que l'on cherche à estimer varie d'un essai à l'autre du fait des différences entre essais (patients étudiés, modalités thérapeutiques, etc.), ces variations étant néanmoins distribuées autour d'une valeur moyenne que cherche à estimer la méta-analyse.

Interaction entre effet du traitement et caractéristiques du patient ou des essais

Quand la quantité d'information accumulée est suffisante, l'estimation de l'effet moyen du traitement peut être complétée par l'étude des facteurs associés à sa variation. Ces analyses sont exploratoires et nécessitent d'être validées par de nouveaux essais.

Ainsi, il est possible d'explorer les variations de l'efficacité du traitement en fonction de diverses caractéristiques des patients comme le sexe, l'âge, l'état général ou diverses caractéristiques de leur tumeur comme le stade, la localisation ou l'histologie, voire d'autres facteurs pronostiques. La méthode habituelle est d'estimer l'effet du traitement dans chaque catégorie de la caractéristique étudiée (par ex., homme et femme) par une analyse stratifiée sur l'essai comme pour l'ana-

Essai	Nb de décès / effectif total CT + RT	CT	O-A	Variance	Hazard ratio (CT + RT : CT)	HR [95% IC]
Copenhague	69/69	74/76	11,2	34,0		1,39 [0,99-1,95]
Sydney	44/45	48/49	-8,2	21,7		0,69 [0,45-1,04]
NCI	46/48	46/49	-8,9	21,3		0,66 [0,43-1,01]
SECSGI	123/153	111/142	-12,1	56,4		0,81 [0,62-1,05]
Londres	59/63	74/75	-7,9	32,5		0,78 [0,56-1,11]
SWOG	43/47	45/56	4,0	21,6		1,20 [0,79-1,83]
SAKK	35/36	32/34	0,6	16,6		1,04 [0,64-1,68]
Uppsala	22/26	31/31	-4,5	12,5		0,70 [0,40-1,21]
CALGB	257/292	128/134	-20,0	75,9		0,77 [0,61-0,96]
ECOG	122/131	121/133	-7,2	59,8		0,89 [0,69-1,14]
Okayama	22/28	27/28	-4,8	12,0		0,67 [0,38-1,18]
SECSG II	116/154	140/168	-10,4	63,1		0,85 [0,66-1,09]
CETCB	14/19	12/17	1,0	6,4		1,17 [0,54-2,54]
Total	**972/1111**	**890/992**	**-67,2**	**433,8**		**0,86 [0,78-0,94]**

Test d'hétérogénéité : χ^2_{12}=16,92 p = 0,15

0,0 0,5 1,0 1,5 2,0

En faveur de CT+RT / En faveur de CT

Effet CT+RT : χ^2 = 10,41 p = 0,001

Figure 1. Exemple de la méta-analyse de la radiothérapie thoracique dans les cancers bronchiques à petites cellules [26] montrant une différence significative (p = 0,001) en faveur de l'association chimiothérapie-radiothérapie (CT + RT) par rapport à la chimiothérapie seule (CT) et ne présentant pas d'hétérogénéité significative (p = 0,15).

Graphique des hazard ratios (rapport des risques instantanés) de chaque essai (carré) avec leur intervalle de confiance à 95 % (ligne horizontale) et du hazard ratio global (pointe du losange et ligne verticale pointillée) avec son intervalle de confiance à 95 % (extrémités du losange). Le trait plein vertical passant par 1 correspond à l'égalité entre les deux traitements. Un hazard ratio à gauche de cette ligne est en faveur du nouveau traitement, à droite en faveur du traitement de référence. La taille de chaque carré est proportionnelle à l'information apportée par l'essai.

O : nombre de décès observés dans le groupe traité ; A : nombre de décès attendus dans le groupe traité en cas d'absence de différence d'efficacité entre le groupe traité et le groupe témoin, calculé selon la méthode du log-rank. Hazard ratio = HR = Exp ((O-A)/Var (O-A)) ; HR global = Exp (Σ (O-A)/ Σ Var (O-A)).

lyse globale et de comparer les résultats obtenus entre les catégories de la variable par un test d'interaction. Ce type d'analyses nécessite de disposer des données individuelles. Si l'ensemble des catégories étudiées d'une caractéristique donnée ne sont pas disponibles pour chaque essai,

l'analyse peut être biaisée. Des méthodes évaluant l'interaction au niveau de chaque essai dans un premier temps, puis réalisant la synthèse des estimations de chaque essai ont été récemment proposées [15].

La méta-analyse peut également être utilisée en rassemblant les résultats d'essais portant sur une question similaire, avec une définition large de la question. Cette vision large des problèmes permet de faire des comparaisons indirectes qui sont source de nouvelles hypothèses à explorer par des essais futurs. Par exemple, dans les cancers ORL, la chimiothérapie dans les formes non métastatiques peut être administrée avant, pendant ou après le traitement locorégional (chirurgie et ou radiothérapie). La méta-analyse correspondante a montré la supériorité de la chimiothérapie concomitante à la radiothérapie, ce qu'a confirmé la comparaison directe [16].

Publication des résultats

La publication d'une méta-analyse doit définir clairement la question étudiée, présenter de façon précise les mots clés et les bases ayant servi à identifier les essais pris en considération et décrire les essais inclus en précisant pour chaque essai : la période d'inclusion, le plan expérimental, la méthode de tirage au sort, les traitements comparés, la population incluse, la qualité du suivi. Ces informations doivent permettre d'évaluer la qualité des essais et d'identifier une éventuelle hétérogénéité clinique. De même, les raisons de l'exclusion de certains essais doivent être explicitées. Les références des essais inclus et exclus doivent être fournies.

Les méthodes utilisées doivent être clairement énoncées. Les représentations graphiques des résultats sont de lecture aisée et doivent être largement utilisées. Une méta-analyse représentant une masse importante de travail, il est utile que les publications soient faites de manière très claire et précise afin qu'une actualisation puisse facilement être réalisée au fur et à mesure que de nouveaux essais sont publiés ou que de nouvelles hypothèses sont envisagées. Le recours à des suppléments disponibles sur le web permet de disposer de l'espace nécessaire pour rapporter l'ensemble des informations nécessaires. Les recommandations pour la publication des méta-analyses ont été récemment actualisées [17]. La qualité des méta-analyses est très variable. Les méta-analyses sur données individuelles ainsi que les méta-analyses sur données publiées réalisées dans le cadre de la *Cochrane Collaboration* sont de meilleure qualité que les autres méta-analyses [18].

Perspectives

Les progrès dans l'identification des essais (registres d'essais), un accès plus facile aux bases de données (progrès technique) et le développement des collaborations internationales devraient conduire à un recours plus fréquent aux méta-analyses sur données individuelles. D'autant qu'une méta-analyse de ce type impliquant moins d'une dizaine d'essais ne nécessite pas de moyens importants. Actuellement, celles-ci sont généralement réservées aux domaines controversés, impliquant des données de survie et des analyses complexes.

Les méta-analyses planifiées de manière prospective, c'est-à-dire avant le début des essais (ou au moins leur analyse), rares actuellement, vont également se développer, car elles permettent plus de souplesse qu'un essai multicentrique international. On peut citer comme exemple de méta-analyse prospective la méta-analyse des *Cholesterol Treatment Trialists'* (CTT) *Collaborators* [19].

Une méta-analyse sur données résumées conduite avec rigueur et cherchant à atteindre la meilleure exhaustivité possible peut permettre d'obtenir des résultats de haute qualité. Elle pourra si besoin être complétée par une méta-analyse sur données individuelles.

À partir des exemples suivants pour lesquels nous recommandons aux lecteurs de lire les articles cités comme première approche, nous souhaitons montrer les possibilités de la méta-analyse au-delà de l'étude standard de l'effet global d'un traitement.

L'étude des causes de mortalité a permis de mettre en évidence la toxicité à long terme de la radiothérapie dans le traitement des patientes atteintes de cancer du sein [20] et l'effet bénéfique de l'aspirine sur la mortalité par cancer [21].

L'étude des modalités de rechute des cancers ORL [16] a permis une meilleure compréhension des différentes modalités d'association de radio-chimiothérapie.

La méta-analyse est maintenant utilisée régulièrement pour valider des critères de jugement de substitution, par exemple la survie sans progression pour la survie globale [22].

Plus récemment, la méta-analyse a été utilisée pour valider l'effet prédictif de marqueurs tumoraux sur l'efficacité de la chimiothérapie dans les cancers bronchiques [23].

Enfin, avec la multiplication des traitements disponibles pour une situation donnée, le recours aux méta-analyses en réseau, qui permet de classer un ensemble de traitements en combinant comparaisons directes et indirectes de ceux-ci, apparaît comme un outil intéressant pour guider la recherche [24, 25].

À retenir

- Deux types de méta-analyses existent : méta-analyse sur données résumées et méta-analyse sur données individuelles (plus longue et coûteuse mais moins exposée aux biais et permettant une analyse plus complète).
- La réalisation d'une méta-analyse de qualité implique une collaboration multidisciplinaire.
- L'exhaustivité des essais pris en considération dans la méta-analyse est très importante. L'obligation de déclarer tout essai clinique en cours dans un registre devrait conduire à faciliter cette étape.

REMERCIEMENTS

Ce travail a bénéficié du soutien de la Ligue nationale contre le cancer.

Références

1. Cucherat M. *Méta-analyse des essais thérapeutiques*. Paris : Masson, collection Évaluation et statistique, 1997, 390 pages.

2. Borenstein M, Hedges LV, Higgins JPT, *et al*. *Introduction to meta-analysis*. Chichester: John Wiley 2009, 450 pages.

3. Egger M, Smith GD, Altman DG, *et al*. *Systematic Reviews in Healthcare*, 2nd edition. London: BMJ Publishing Group, 2001, 487 pages.

4. Sutton AJ, Abrams KR, Jones DR, *et al. Methods for meta-analysis in medical research.* Chichester: John Wiley, 2000, 346 pages.

5. Parmar MKB, Torri V, Stewart L. Extracting summary statistics to perform meta-analyses of the published literature for survival endpoints. *Stat Med* 1998 ; 17 : 2815-34.

6. Michiels S, Piedbois P, Burdett S, *et al.* Meta-Analysis when only the median survival times are known: A comparison with individual patient data results. *Int J Technol Assess Health Care* 2005 ; 21 : 119-25.

7. Stewart LA, Clarke MJ on behalf of the Cochrane Working Group on meta-analyses using individual patient data. Practical methodology of meta-analyses (overviews) using updated individual patients data. *Stat Med* 1995 ; 14 : 2057-79.

8. Jadad AR, Moore RA, Carroll D, *et al.* Assessing the quality of reports of randomized clinical trials: Is blinding necessary? *Control Clin Trials* 1996 ; 17 : 1-12.

9. Jüni P, Witschi A, Bloch R, *et al*. The hazard of scoring the quality of clinical trials for meta-analysis. *JAMA* 1999 ; 282 : 1054-60.

10. Berlin JA, Laird NM, Sacks HS, *et al*. A comparison of statistical methods for combining event rates from clinical trials. *Stat Med* 1889 ; 8 : 141-51.

11. Der Simonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clin Trials* 1986 ; 7 : 177-88.

12. Galbraith RF. A note on graphical presentation of estimated odds ratios from several clinical trials. *Stat Med* 1988 ; 7 : 889-94.

13. Baujat B, Mahé C, Pignon JP, *et al*. A graphical method for exploring heterogeneity in meta-analysis: Application to a meta-analysis of 65 trials. *Stat Med* 2002 ; 21 : 2641-52.

14. Yusuf S, Wittes J, Probstfield J, *et al.* Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *JAMA* 1991 ; 266 : 93-8.

15. Fisher DJ, Copas AJ, Tierney JF, *et al*. A critical review of methods for the assessment of patient-level interactions in individual patient data meta-analysis of randomized trials, and guidance for practitioners. *J Clin Epidemiol* 2011 : 2011 ; 64 (9) : 949-67.

16. Pignon JP, Le Maître A, Maillard E, *et al.* Meta-analysis of chemotherapy in head & neck cancer (MACH-NC): An update on 93 randomized trials and 17 346 patients. *Radiother Oncol* 2009 ; 92 : 4-14.

17. Moher D, Alessandro Liberati A, Tetzlaff J, *et al*. Preferred Reporting Items for Systematic Reviews and Meta-Analyses : The PRISMA Statement. *Ann Intern Med* 2009 ; 151 : 1-6.

18. Brignone M, Aupérin A, Colombet I, *et al*. Critical review of meta-analyses analyzing lung cancer treatment. 1st European Lung Cancer Conference, Geneva, April 2008. *J Thorax Oncol* 2008 ; 3 (suppl. 1) : S47.

19. Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90 056 participants in 14 randomised trials of statins. *Lancet* 2005 ; 366 : 1267-78.

20. Early Breast Cancer Trialists Collaborative Group. Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: An overview of the randomized trials. *Lancet* 2000 ; 355 : 1757-70.

21. Rothwell PM, Fowkes FGR, Belch JFF, *et al*. Effect of daily aspirin on long-term risk of death due to cancer: Analysis of individual patient data from randomised trials. *Lancet* 2011 ; 377 : 31-41.

22. Michiels S, Le Maître A, Buyse M, *et al.* ; on behalf of the MARCH and MACH-NC Collaborative Groups. Surrogate endpoints for overall survival in locally advanced head and neck cancer: Meta-analyses of individual patient data. *Lancet Oncol* 2009 ; 10 : 341-50.

23. Reiman T, Lai R, Veillard AS, *et al.* ; for the LACE-Bio Group. Cross-validation study of Class III Beta-Tubulin as a predictive marker for benefit from adjuvant chemotherapy in resected non-small cell lung cancer: analysis of four randomized trials. *Ann Oncol* 2011 april 6 [ahead of print].

24. Blanchard P, Hill C, Guihenneuc-Jouyaux C, *et al.* ; on behalf of the MACH-NC and MARCH Collaborative Groups. Mixed treatment comparison meta-analysis of altered fractionated radiotherapy and chemotherapy in head and neck cancer. *J Clin Epidemiol* 2011 : 2011 ; 64 (9) : 985-92.

25. Lumley T. Network meta-analysis for indirect treatment comparisons. *Stat Med* 2002 ; 21 : 2313-24.

26. Pignon JP, Arriagada R, Ihde DC, *et al.* A meta-analysis of thoracic radiotherapy for small-cell lung cancer. *N Engl J Med* 1992 ; 327 : 1618-24.

Analyse statistique des données d'expression de gènes issues de puces à ADN

É. Gravier, F. Valet

La technologie des puces à ADN, née dans les années 1990, permet la mesure simultanée du niveau d'expression de dizaines de milliers de gènes d'un échantillon biologique. L'utilisation de cette technologie n'a cessé de croître ces dernières années et les puces à ADN occupent désormais une place prépondérante dans la recherche clinique en cancérologie. Deux signatures basées sur le profil d'expression de gènes ont été proposées comme facteurs pronostiques dans le cancer du sein et sont actuellement évaluées dans deux grands essais cliniques en Europe et aux États-Unis [1, 2]. Les problématiques rencontrées dans le domaine de la génomique sont principalement l'amélioration de la classification des tumeurs et l'identification de mécanismes biologiques associés au devenir clinique des malades ou à une réponse thérapeutique. La spécificité majeure des données de puces à ADN est le grand nombre de variables mesurées (produit d'expression des gènes) relativement au petit nombre d'individus observés (les tumeurs). L'objectif est donc d'exploiter et d'interpréter au mieux l'information issue de données d'expression de gènes en prenant en compte ce « fléau de la dimension ». Ce chapitre présente les méthodes statistiques les plus communément utilisées pour l'analyse de ce type de données.

Le principe des puces à ADN

Schématiquement, une puce à ADN est un support solide (plastique ou verre par ex.) sur lequel sont régulièrement répartis de très nombreux fragments d'ADN, appelés « sondes ». La séquence de bases de chaque sonde est connue et est caractéristique de la région d'un gène. Chaque sonde a la capacité de se lier spécifiquement à la séquence d'ADN qui lui est complémentaire (dite séquence « cible »). Les ARN messagers sont extraits de l'échantillon à analyser puis transformés en ARN complémentaires sur lesquels sont fixées des molécules de biotine (qui permettront ensuite de réaliser un marquage fluorescent). Ces séquences d'ARN complémentaires sont hybridées aux sondes présentes sur la puce par complémentarité de séquences. Un marqueur fluorescent ayant la propriété de se fixer aux molécules de biotine est alors introduit, laissant apparaître un spot lumineux à l'endroit de l'hybridation. L'image de la puce est ensuite obtenue grâce à un scanner. L'analyse de cette image permet finalement de quantifier l'intensité de chaque spot, proportionnelle au niveau d'expression du gène dans l'échantillon analysé *(figure 1)*.

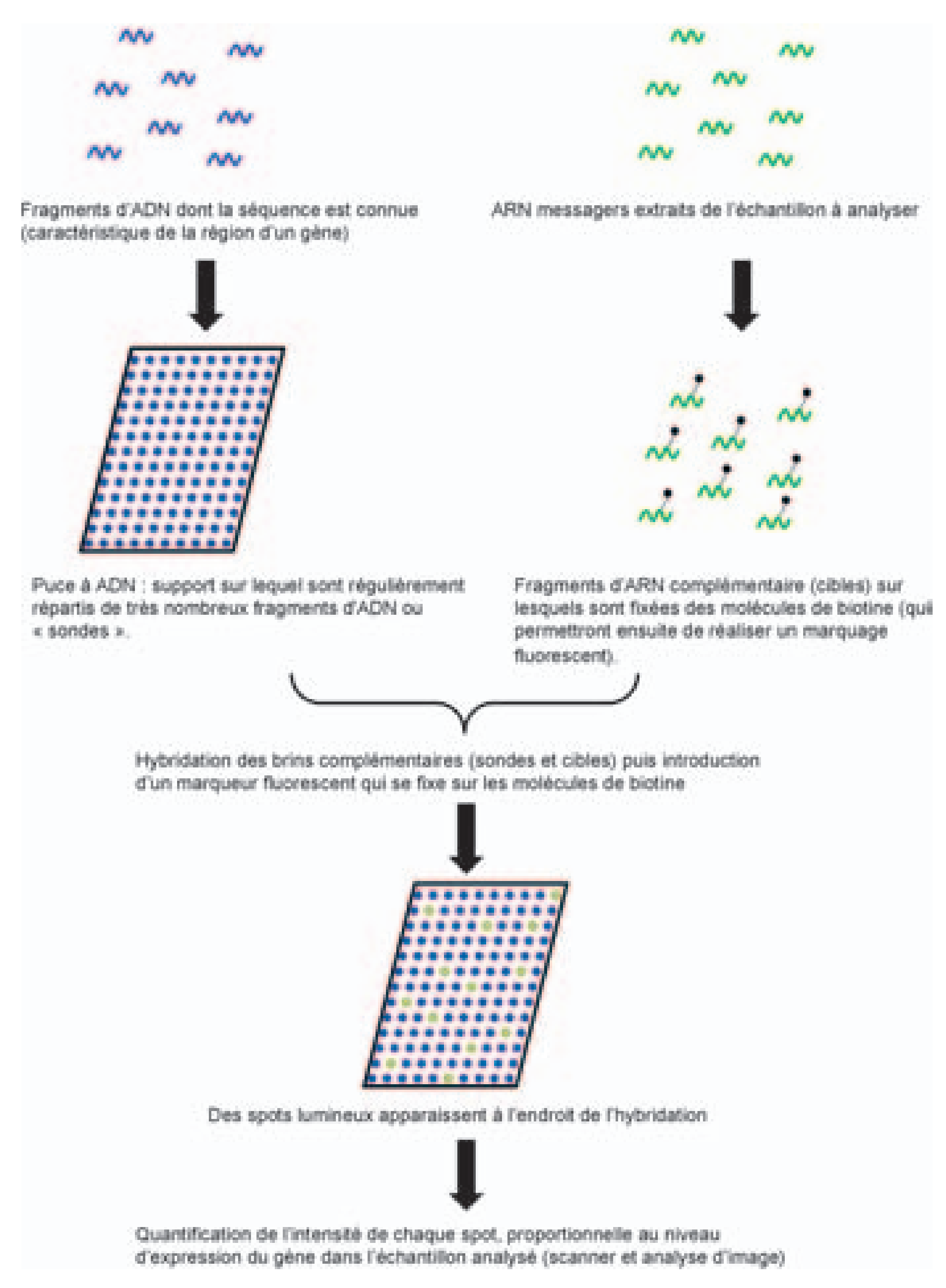


Figure 1. Principe des puces à ADN permettant la mesure de l'expression des gènes.

Différentes technologies de puces à ADN « expression » existent (par ex., puces commercialisées par les sociétés Affymetrix ou Agilent). En dehors des méthodes de normalisation qui sont spécifiques d'une technologie, les méthodes statistiques présentées dans ce chapitre s'appliquent aux différents types de puces.

Le prétraitement des données

La normalisation

L'obtention de mesures d'expression de gènes est un processus élaboré qui s'accompagne de nombreuses sources de variabilité techniques et expérimentales. Il s'agit, par exemple, de biais en lien avec le processus de *spotting* (dépôt des sondes sur la puce), l'extraction et le marquage des cibles, l'hybridation non spécifique (hybridation de la sonde avec des séquences différentes de la séquence cible), les erreurs de mesure introduites par le scanner et l'analyse d'image, les conditions d'expérimentation (par ex. la température)… Ces différentes sources de biais se confondent avec la variabilité biologique étudiée, voilà pourquoi il est primordial de s'en affranchir avant toute analyse : c'est l'objectif de la normalisation qui permet ainsi de rendre les puces comparables. À chaque technologie de puce est associée une méthode de normalisation. Dans le cas de puces de type Affymetrix (technologie dominante à l'heure actuelle), les méthodes de normalisation les plus utilisées sont MAS 5.0 [3], RMA [4] et GCRMA [5].

Le filtrage des données

La procédure de filtrage a pour but de supprimer, avant toute analyse, les gènes dont l'expression ne varie pas ou peu, dans le jeu de puces étudié. La justification de cette étape de prétraitement des données est d'abord biologique : on supprime les gènes qui ne présentent *a priori* aucune variabilité biologique. L'argument est également méthodologique : cette étape permet de réduire le nombre de tests lors de l'étape de l'analyse différentielle. Les procédures de filtrage classiques éliminent les gènes dont l'écart-type ou l'intervalle inter-quartile est inférieur à un seuil donné. Le choix du seuil est cependant arbitraire et peut parfois fortement influencer les analyses ultérieures.

Les analyses non supervisées

Les analyses non supervisées ont pour but de découvrir des sous-groupes de tumeurs (ou de gènes) aux profils d'expression similaires, indépendamment de toute connaissance clinique *a priori*. Ce sont donc des techniques à visée purement descriptive. Ces méthodes aident à la détection de tumeurs au profil atypique ou à l'identification de biais et sont ainsi utilisées afin de contrôler la qualité des données. Elles permettent également d'identifier des sous-groupes de tumeurs biologiquement homogènes au sein desquels il peut être intéressant de rechercher des facteurs

pronostiques ou prédictifs. Les analyses non supervisées les plus communément rencontrées en génomique sont l'analyse en composantes principales (ACP) ou les analyses en *clusters* (kmeans, classification ascendante hiérarchique). Nous développons dans cette partie la classification ascendante hiérarchique.

Le principe de la classification ascendante hiérarchique

Considérons que chaque individu constitue une classe (ou *cluster singleton*). La première étape consiste à regrouper au sein d'un *cluster* à deux éléments les deux individus les plus « proches ». Cette notion de proximité suppose donc la définition d'une distance entre deux individus, que nous aborderons plus loin. Les étapes suivantes regroupent successivement les *clusters* les plus « proches » jusqu'à l'obtention d'un unique *cluster* regroupant tous les individus. Ces étapes de regroupements successifs de *clusters* supposent la définition d'une seconde distance (appelé critère d'agrégation), mesurant la proximité non plus entre deux individus mais entre deux *clusters* d'individus.

Distances entre individus

Soient deux vecteurs x et y de dimension p tels que $x = (x_1, ..., x_p)$ et $y = (y_1, ..., y_p)$.

Les distances entre x et y les plus communément utilisées en génomique sont :

- la distance euclidienne :

$$d_E(x,y) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

- la distance du coefficient de corrélation de Pearson :

$$d_\rho(x,y) = \frac{(1-\rho(x,y))}{2}$$

où ρ est le coefficient de corrélation de Pearson ;

- la distance du coefficient de corrélation de Pearson absolu :

$$d_{|\rho|}(x,y) = 1-|\rho(x,y)|$$

Généralement, il n'est pas pertinent de grouper des tumeurs aux profils d'expression anti-corrélés, voilà pourquoi d_ρ est la distance plutôt utilisée dans la classification de tumeurs. En ce qui concerne la classification des gènes, il est intéressant de pouvoir grouper des gènes appartenant à la même voie de signalisation (s'activant ou s'inhibant mutuellement). Dans ce cadre, on

utilisera plutôt $d_{|\rho|}$. À noter qu'il peut être utile de remplacer le coefficient de corrélation de Pearson par le coefficient de corrélation de Spearman (notamment dans les cas de distributions non symétriques ou de faible nombre d'individus).

Critères d'agrégation

Soient A et B deux *clusters* de cardinaux respectifs n_A et n_B. Le problème est de définir une distance *d(A,B)* entre ces deux *clusters* à partir de la distance *d* entre individus (par ex. d_E, d_ρ, ou $d_{|\rho|}$). Les critères d'agrégation les plus communément utilisés en génomique sont :

- le saut moyen :

$$d(A,B) = \frac{1}{n_A \times n_B} \sum_{x \in A, y \in B} d(x,y)$$

- le saut de Ward

$$d(A,B) = \frac{n_A \times n_B}{n_A + n_B} d^2(g_A, g_B)$$

où g_A et g_B sont les barycentres respectifs de A et de B.

Le saut de Ward est le critère d'agrégation le plus couramment utilisé lorsque la distance entre individus utilisée est la distance euclidienne. Dans ce cas, il représente la perte de variance inter-classe générée par le regroupement des deux *clusters* A et B. Cependant, ce critère peut également être utilisé en association avec d'autres distances, même s'il n'est alors plus interprétable. Dans le cas de distances basées sur la corrélation, les critères du saut de Ward et du saut moyen sont en pratique utilisés.

Représentation graphique

La classification hiérarchique ascendante produit un arbre binaire de classification (ou dendrogramme). Cet arbre représente graphiquement les agrégations successives depuis les *clusters singletons* représentant un individu (« feuilles » de l'arbre) jusqu'au *cluster* regroupant l'ensemble des individus (« racine » de l'arbre). La hauteur de la branche qui unit deux *clusters* est proportionnelle à la distance entre ces *clusters* avant regroupement.

En génomique, il est courant de représenter sur un même graphique les résultats de classification des tumeurs et des gènes. Les valeurs d'expression sont présentées sous forme matricielle : en ligne les gènes (ordonnés par le « dendrogramme des gènes ») et en colonne les tumeurs (ordonnées par le « dendrogramme des tumeurs »). La cellule correspondant à l'intersection d'une ligne et d'une colonne est colorée selon la valeur d'expression du gène dans la tumeur considérée. Cette représentation, appelée *heatmap* [6], permet non seulement de visualiser les similitudes d'expression entre gènes et entre tumeurs mais également de déterminer les groupes de gènes qui ont influencé la classification des tumeurs.

Un exemple : la classification moléculaire des cancers du sein

La classification hiérarchique des profils d'expression de tumeurs du sein a permis d'identifier cinq sous-types de tumeurs aux caractéristiques histologiques distinctes et dont les pronostics se sont révélés différents [7-9]. La *figure 2* présente les résultats de cette classification ainsi que les valeurs d'expression *(heatmap)* des cinq groupes de gènes caractéristiques de chaque sous-type (six sous-types sont présentés sur la figure mais il a été montré ultérieurement que le sous-type luminal C était marginal et non robuste, voilà pourquoi on parle plutôt de cinq sous-types).

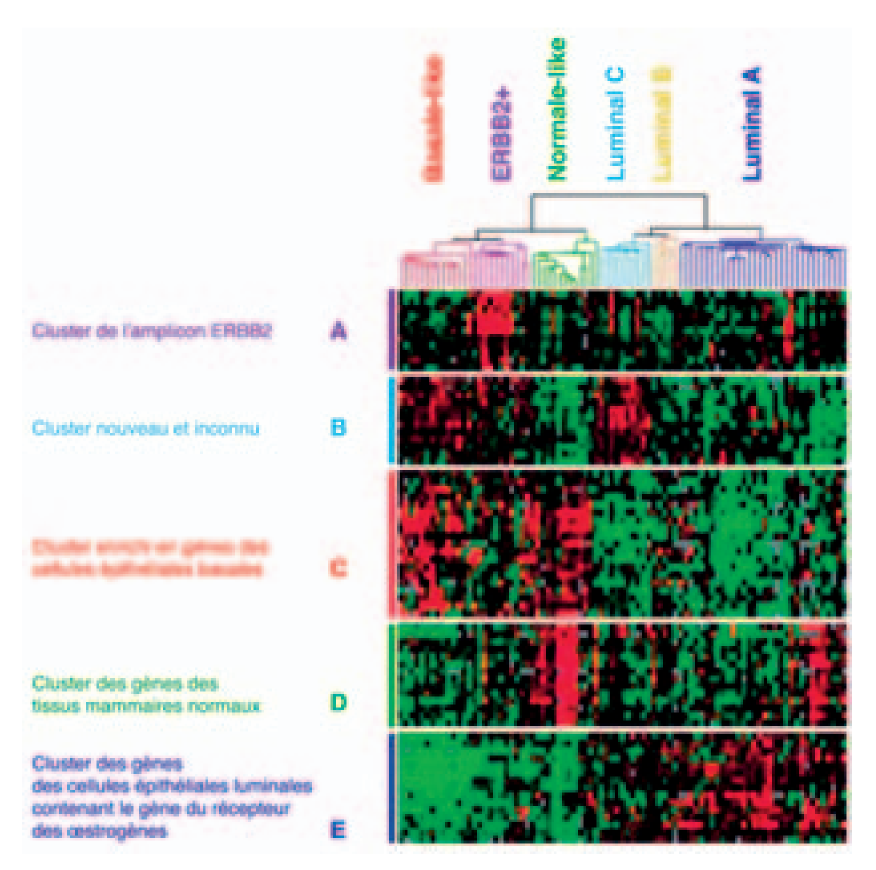


Figure 2. *Clustering* hiérarchique de 85 tumeurs du sein basé sur le profil d'expression de 427 gènes (distance de corrélation de Pearson et critère d'agrégation du saut moyen).
Les tumeurs (en colonne) ont été classées en cinq sous-types. Les valeurs d'expression des cinq groupes de gènes caractéristiques de chaque sous-type sont présentées sous forme de heatmap. *En vert : les valeurs d'expression faibles, en noir : les valeurs d'expression médianes et en rouge : les valeurs d'expression fortes.*

Les analyses supervisées

Les analyses supervisées des données de puces à ADN sont de deux types. L'analyse différentielle a pour objectif d'identifier les gènes dont l'expression diffère significativement entre deux ou plusieurs classes de patients (par ex. deux groupes de patients aux devenirs cliniques différents). La classification supervisée permet quant à elle de prédire la classe d'une tumeur sur la base de son profil d'expression. Par souci de simplicité (et également car il s'agit du cas le plus fréquent), nous nous placerons dans la situation où la variable à expliquer est une variable binaire.

L'analyse différentielle

Introduisons ici quelques notations.

Pour chaque gène *i*, soient :
- H_{0i} l'hypothèse nulle à tester : « pas de différence d'expression du gène *i* entre les deux classes » ;
- H_{1i} l'hypothèse alternative : « différence d'expression du gène *i* entre les deux classes » ;
- T_i la statistique de test mesurant la différence d'expression du gène *i* entre les deux classes ;
- p_i la *p-value* associée à T_i, mesurant pour le gène le risque d'être faux positif.

Supposons dans cette partie que l'on teste m hypothèses nulles parmi lesquelles m_0 sont vraies. Soit donc $\pi_0 = m_0/m$ la proportion d'hypothèses nulles vraies parmi les m hypothèses testées. Notons α le risque de première espèce associé à chacun de ces tests.

Le problème des tests multiples

Dans le cas d'un grand nombre d'hypothèses à tester (des dizaines de milliers en génomique), il faut être conscient que le risque de faux positifs est accru. Pour illustrer notre propos, supposons par exemple que l'on teste $m = 50\,000$ hypothèses nulles au niveau $\alpha = 0{,}05$ et que toutes soient vraies ($m_0 = m$). Alors on s'attend à rejeter (à tort !) en moyenne 2 500 (= 0,05*50 000) de ces hypothèses. Il devient alors évident que, dans un contexte de tests multiples, le contrôle du risque individuel de première espèce α n'est plus suffisant.

Afin de pallier ce problème, Benjamini et Hochberg introduisent en 1995 une nouvelle mesure du risque : le *False Discovery Rate* (FDR) [10]. Ils définissent ce taux de « fausses découvertes » comme l'espérance de la proportion de faux positifs parmi les gènes déclarés différentiellement exprimés. C'est la mesure du risque la plus utilisée dans l'analyse de données de génomique. Les p_i sont ainsi « corrigées » en p_i^* pour refléter désormais le niveau de FDR. Sélectionner les gènes dont $p_i^* \leq 0{,}05$ permet de contrôler le FDR au niveau 5 %. Les deux procédures d'estimation du FDR les plus fréquemment utilisées sont présentées ci-dessous.

La procédure de Benjamini et Hochberg

Soient $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$ les *p-values* obtenues, ordonnées par ordre croissant.

Soient $H_{0(1)}$, $H_{0(2)}$, ..., $H_{0(m)}$ les hypothèses nulles associées.

Soit k vérifiant :

$$k = \underset{i}{\arg\max}\left(p_{(i)} \leq \frac{i}{m}\alpha \right)$$

Alors, sous l'hypothèse d'indépendance des statistiques de tests, rejeter les hypothèses $H_{0(i)}$, $i = 1, ..., k$ contrôle le FDR au niveau $m_0\alpha/m \leq \alpha$. À noter que cette méthode ne permet pas l'estimation de m_0, supposé égal à m.

Significance Analysis of Microarrays (SAM)

Cette méthode d'estimation du FDR [11] utilise des permutations des données. Elle présente l'avantage de prendre en compte la structure de dépendance des gènes, d'éviter de faire des hypothèses sur leur distribution, mais également d'estimer m_0.

Une statistique de test est calculée pour chaque gène. L'estimation des distributions des statistiques sous l'hypothèse qu'aucun gène n'est différentiellement exprimé est ensuite réalisée en permutant aléatoirement les valeurs de la variable à expliquer (attribution d'une valeur de classe existante pour chacune des tumeurs). La valeur de la statistique observée de chaque gène est alors comparée à sa valeur attendue sous l'hypothèse nulle. Un gène est déclaré différentiellement exprimé si la différence entre valeur observée et attendue (en valeur absolue) dépasse un certain seuil Δ. Notons par la suite L, la liste de gènes déclarés différentiellement exprimés au seuil Δ. Le nombre de faux positifs associé à ce seuil sous l'hypothèse qu'aucun gène n'est différentiellement exprimé ($n_{\mathrm{FP}|\pi 0 = 1}$) est estimé en comptant le nombre médian de gènes qui sont déclarés différentiellement exprimés sur les données permutées. Le nombre de faux positifs attendu dans L est estimé par $n_{\mathrm{FP}} = \pi_0 \times n_{\mathrm{FP}|\pi 0 = 1}$ (voir [12] pour le détail de l'estimation de π_0). Le FDR est finalement calculé comme le quotient de n_{FP} sur le nombre de gènes de L.

Interprétation biologique des résultats d'analyse différentielle

Une fois détectés les gènes différentiellement exprimés, il s'agit d'identifier les mécanismes biologiques qui leur sont associés. L'approche la plus classique est de rechercher parmi des groupes de gènes impliqués dans les mêmes processus biologiques (par ex. groupe de gènes associé à l'apoptose, la prolifération cellulaire, l'activité des protéines kinase) ceux qui sont surreprésentés au sein de la liste L. On dit que l'on recherche si cette liste de gènes est « enrichie » en annotations. Pour chaque annotation, le principe est de tester si les fréquences de l'annotation au sein de la sélection et au sein de l'ensemble de la puce sont égales. En pratique, les annotations sont fournies par une base de données telle que *Gene Ontology* [13] ou KEGG [14]. Le degré d'enrichissement de L pour une annotation donnée est évalué en utilisant la loi hypergéométrique. Des milliers d'annotations sont testées, voilà pourquoi il est également nécessaire d'appliquer une procédure de contrôle du FDR.

Les résultats de ce type d'analyse dépendent du seuil de significativité choisi pour définir la liste *L* de gènes sélectionnés. Dans le but de s'affranchir du choix de ce seuil, différentes méthodes ont été développées [15, 16] basées sur la comparaison de la distribution des statistiques de tests des gènes partageant l'annotation (sur l'ensemble de la puce) à celle des gènes ne la partageant pas. Enfin, sont également proposées des approches testant directement la liaison entre le groupe de gènes et la variable à expliquer [17].

La limitation principale de toutes ces approches est qu'elles se basent sur des groupes de gènes connus *a priori*, ce qui laisse peu de place à la découverte de nouveaux groupes de gènes biologiquement pertinents.

La classification supervisée

L'objectif est de prédire la classe d'une tumeur (par ex. bon ou mauvais devenir clinique) sur la base de son profil d'expression. La stratégie d'analyse comprend généralement deux étapes. L'étape d'apprentissage consiste à construire la règle de décision (ou signature). L'étape de validation a pour but d'évaluer ses performances.

Construction d'une signature

Les données de puces à ADN sont caractérisées par un petit nombre d'individus (quelques centaines) et un très grand nombre de variables (plusieurs centaines de milliers). Ce « fléau de la dimension » souvent associé à une situation de colinéarité rend les méthodes de régression classiques difficilement applicables (en particulier forte variance des estimateurs des moindres carrés) et induit un risque majeur de surajustement [18]. La solution consiste à réduire la dimension des données. Trois types d'approches peuvent être utilisés. Les méthodes de sélection classiques (de type ascendante ou *stepwise*) permettent la sélection de quelques gènes seulement. Ces méthodes posent en particulier le problème de robustesse (la signature dépend fortement du jeu de données sur laquelle elle est apprise). Les méthodes de projection permettent quant à elles de déterminer un petit nombre de variables indépendantes (composantes) construites comme combinaisons linéaires des gènes. Il s'agit, par exemple, de la régression sur composantes principales ou de la régression *Partial Least Square* [19].

Enfin, les méthodes les plus largement utilisées sont les méthodes de régression avec pénalisation de type *ridge regression* [20], lasso [21] ou elasticnet [22]. Elles permettent de contourner le problème de colinéarité entre les variables explicatives en ajoutant une contrainte sur la norme des coefficients de régression. Les méthodes lasso et elasticnet présentent par ailleurs l'avantage de permettre la sélection de variables.

Validation

L'idéal est d'appliquer la signature construite sur le jeu d'apprentissage à un jeu de données totalement indépendant (jeu de validation). Les performances de la règle de décision sont estimées sur ce jeu en calculant le taux de mauvaises classifications, la sensibilité et la spécificité

ainsi que leurs intervalles de confiance. Dans le cas de données de génomique, où (rappelons-le) le nombre d'individus est faible, tous les individus participent souvent à l'apprentissage et il est donc difficile de disposer d'un jeu de données indépendant. Dans ce cas-là, la signature S est construite sur l'ensemble des *n* individus puis validée par « validation croisée » [23]. Le principe est de partitionner l'ensemble du jeu disponible en *k* groupes d'effectifs et de caractéristiques clinico-pathologiques comparables. La règle de décision est ensuite construite sur les individus des *k-1* premiers groupes (qui jouent donc le rôle du jeu d'apprentissage), et ses performances sont évaluées sur le dernier groupe *k* (qui joue le rôle de jeu de validation). La procédure est réitérée *k* fois de sorte que chaque groupe joue le rôle de jeu de validation. Les performances sont finalement estimées en moyennant les *k* taux de mauvaises classifications obtenus. Cette méthode de rééchantillonnage porte le nom de *k-folds* (par ex. *ten-folds*, *five-folds*) et dans le cas où $k = n$, elle est nommée *leave one out*. Lorsque *k* augmente (c'est-à-dire lorsque le nombre d'individus qui contribuent à la phase d'apprentissage augmente), le biais de l'estimation des performances diminue, mais sa variance augmente. En pratique, le choix de *k* dépend donc du compromis biais/variance que l'on souhaite. À noter que les *k* signatures générées lors de la procédure de validation croisée sont potentiellement différentes de *S*, car construites sur un ensemble d'individus différent. Les techniques de validation croisée valident donc la méthode de construction de *S*, plutôt que *S* elle-même.

Conclusion

La haute dimensionnalité des données de puces à ADN pose des problèmes méthodologiques à chaque étape de l'analyse statistique. De nombreuses approches ont été développées ces dernières années pour prendre en compte cette spécificité, mais aucun consensus n'a vraiment été établi concernant le choix d'une méthode plutôt qu'une autre. Ce chapitre présente un éventail des méthodes les plus communément utilisées, de l'étape du prétraitement des données à celle de la construction/validation d'une signature.

Quelle que soit l'étape d'analyse, le choix de la stratégie statistique doit être orienté par les discussions avec biologistes et/ou cliniciens, qui se révèlent en particulier indispensables lors de la mise en place du plan expérimental, de l'interprétation biologique et de l'évaluation de la pertinence clinique des résultats.

D'autres types de puces à ADN permettent l'analyse du nombre de copies d'ADN et/ou du déséquilibre allélique (puces *Comparative Genomic Hybridization* et *Single Nucleotide Polymorphism*). L'analyse de ce type de données nécessite la prise en compte de l'information de proximité des variables sur le génome, afin de mettre en évidence des régions chromosomiques altérées et/ou présentant un déséquilibre allélique. Des puces épigénome ou protéome existent également, ces dernières ayant une importance particulière, puisque représentant l'étape finale et effective de l'expression du matériel génétique.

Alors que les problèmes méthodologiques liés aux données de puces à ADN ne sont pas tous résolus, l'utilisation de ce type de données semble être déjà révolue. En effet, l'ère ambitieuse du séquençage massif lui a succédé [24], augurant ainsi de nouveaux défis méthodologiques...

Références

1. Van 'T Veer LJ, Dai H, van de Vijver MJ, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002 ; 415 (6871) : 530-6.

2. Paik S, Shak S, Tang G, *et al*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004 ; 351 (27) : 2817-26.

3. Affymetrix. *Statistical Algorithms Description Document*. Affymetrix, 2002: http://www.affymetrix.com/support/technical/whitepapers.affx.

4. Irizarry RA, Hobbs B, Collin F, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003 ; 4 (2) : 249-64.

5. Wu Z, Irizarry RA, Gentleman R, *et al.* A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *J Am Stat Assoc* 2004 ; 99 (468) : 909-17.

6. Eisen MB, Spellman PT, Brown PO, *et al*. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998 ; 95 (25) : 14863-8.

7. Perou CM, Sorlie T, Eisen MB, *et al.* Molecular portraits of human breast tumours. *Nature* 2000 ; 406 (6797) : 747-52.

8. Sorlie T, Perou CM, Tibshirani R, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001 ; 98 (19) : 10869-74.

9. Sorlie T, Tibshirani R, Parker J, *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci USA* 2003 ; 100 (14) : 8418-23.

10. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B* 1995 ; 57 (1) : 289-300.

11. Tusher VG, Tibshirani R, Chu G, *et al.* Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 2001 ; 98 (9) : 5116-21.

12. Chu G, Narasimhan B, Tibshirani R, *et al.* SAM "Significance Analysis of Microarrays" Users guide and technical document.

13. Harris MA, Clark J, Ireland A, *et al.* The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004 ; 32 (Database issue) : D258-61.

14. Ogata H, Goto S, Sato K, *et al*. KEGG : Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999 ; 27 (1) : 29-34.

15. Subramanian A, Tamayo P, Mostha VK, *et al*. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* 2005 ; 102 (43) : 15545-50.

16. Efron B, Tibshirani R. On testing the significance of sets of genes. *Ann Appl Stat* 2007 ; 1 (1) : 107-29.

17. Goeman JJ, Van De Geer SA, De Kort F, *et al.* A global test for groups of genes: Testing association with a clinical outcome. *Bioinformatics* 2004 ; 20 (1) : 93-9.

18. Hastie T, Tibshirani R, Friedman J, *et al. Linear Methods for regression. The elements of statistical learning: Data mining, inference and prediction.* New York : Springer, 2001.

19. Tenenhaus M. *La régression PLS : théorie et pratique*. Paris : Éditions Technip, 1998.

20. Hoerl AE, Kennard RW. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics* 1970 ; 12 (1) : 55-68.

21. Tibshirani R. Regression Shrinkage and Selection *via* the Lasso. *J R Statist Soc B* 1996 ; 58 (1) : 267-88.

22. Zou H, Hastie T. Regularization and variable selection *via* the elastic net. *J R Statist Soc B* 2005 ; 67 (2) : 301-20.

23. Molinaro AM, Simon R, Pfeiffer RM, *et al*. Prediction error estimation: A comparison of resampling methods. *Bioinformatics* 2005 ; 21 (15) : 3301-7.

24. Ledford H. The death of microarrays? *Nature* 2008 ; 455 : 847.

Partie V

Les différentes phases d'un essai thérapeutique

Planification d'un essai de phase I

E. Chamorey, J. Gal, N. Houédé, X. Paoletti

L'essai clinique de phase I est la première étape d'évaluation d'un nouvel agent thérapeutique chez l'homme. Pour les traitements présentant un index thérapeutique large, ces études sont conduites chez des volontaires sains. Dans le cas des traitements de cancérologie, les essais de phase I sont menés sur le volontaire atteint de cancer, en échec thérapeutique des traitements validés. L'objectif principal est de définir le plus rapidement et précisément possible la dose recommandée (DR) pour les futurs essais de phase II. Il conviendra d'estimer les toxicités dose limitantes (TDL) qui sont les différents types de toxicité limitant l'augmentation de dose, ainsi que la dose maximale tolérée (DMT) qui est la dose entraînant des TDL intolérables pour une proportion donnée de patients traités (le plus souvent 1/3) [1, 2]. On distingue les phases I de première administration chez l'homme pour lesquelles, en règle générale, une seule molécule est administrée des phases I menées sur des traitements déjà utilisés sur l'homme, évaluant une nouvelle forme galénique, un nouveau schéma d'administration ou une association de plusieurs molécules. La planification d'une étude de phase I de cancérologie s'articule autour de trois étapes [2] :

– sélection de la dose initiale entraînant une toxicité présumée acceptable en fonction des résultats des études précliniques ou cliniques ;
– choix des posologies pour les paliers de doses successifs ;
– choix d'un schéma d'escalade de dose.

Les patients sont inclus selon des paliers de doses croissants successifs depuis la dose initiale. Les inclusions des patients procèdent séquentiellement et, généralement, l'essai s'interrompt lorsqu'une proportion définie de malades ayant reçu la même dose développe des TDL. La méthodologie statistique d'une étude de phase I devra permettre de ne pas sous-traiter ou sur-traiter les patients, de minimiser le nombre total de patients inclus et de recueillir des informations objectives et précises sur la DMT, la DR et les TDL.

Remarque

Il existe deux définitions différentes pour la DMT. Les Anglo-Saxons appellent MTD *(Maximum Tolerated Dose)* la dose la plus élevée n'entraînant pas d'effets indésirables inacceptables. Elle correspond en Europe à la dose recommandée (DR), ce qui peut induire certaines confusions à la lecture d'articles scientifiques ou lors d'utilisation d'application informatique spécialisée. Les Européens appellent DMT (dose maximale tolérée) la dose à laquelle on observe une fréquence intolérable de toxicité dose limitante (TDL), en règle générale plus de 30 %, c'est la

posologie à laquelle on arrête l'escalade de dose. Pour éviter ce risque de confusion, les Anglo-Saxons ont introduit le terme de MAD *(Maximum Administered Dose)* qui correspondrait à la DMT européenne ; ce terme est peu usité.

Choix de la dose initiale

Pour les essais de première administration chez l'homme, le choix de la dose initiale est basé sur les résultats des études précliniques. De nombreux auteurs ont étudié les relations entre la toxicologie animale et l'estimation de la dose initiale administrable chez l'homme ainsi que les corrélations entre les surfaces corporelles des espèces étudiées [3, 4]. Classiquement, pour les agents anticancéreux, on utilise la dose correspondant à 1/10e de la dose létale chez 10 % des souris non atteintes de cancer (MELD10 pour *Murine Equivalent Letal Dose* 10). Cette correspondance n'est valable que pour la même voie d'administration et à condition que les doses soient normalisées en surfaces corporelles (mg/m^2) [4].

Pour les études menées sur des traitements déjà utilisés sur l'homme, le choix de la dose initiale est extrapolé en considérant les résultats des études cliniques précédentes et les propriétés pharmacologiques des molécules testées.

Pour les associations de molécules, on pourra extrapoler la dose initiale en fonction des données de chaque molécule prises indépendamment ou sur des études d'associations menées sur des composés de pharmacologie similaire.

Remarque

L'administration de doses exactes selon la surface corporelle du patient n'est possible que dans le cas des molécules d'administration parentérale. Les posologies sont souvent dictées par le dosage de la spécialité pharmaceutique.

Choix des paliers de dose

Après avoir sélectionné la dose initiale, il convient de définir les paliers de dose qui seront testés. La méthode la plus utilisée pour sélectionner ces paliers de dose est dérivée de la suite de « Fibonacci » (1, 1, 2, 3, 5, 8, 13, 21, 34...), chaque chiffre correspondant à la somme des deux chiffres précédents. En pratique, la séquence de Fibonacci a été modifiée : après normalisation sur la dose initiale, les facteurs d'augmentation de dose deviennent : 1 / 2 / 3,3 / 5 / 7 / 9,3 / 12,4... Cette modification permet d'accroître les doses rapidement pour les premiers paliers puis de s'orienter vers une augmentation constante de 33 % après un certain seuil. Il existe d'autres méthodes pour remplacer cette règle qui n'a pas de fondement pharmacologique. En particulier, des incréments de dose reposant sur les résultats observés (augmenter plus fortement la dose si des toxicités faibles ont été observées et ralentir l'escalade en cas de toxicités modérées ou sévères) paraissent

plus efficaces. Pour les molécules ayant déjà été étudiées chez l'homme ou pour les associations de molécules, le rapport entre les paliers de doses successives est généralement fixé à ≈ 1,3. Ce rapport peut être modulé selon le profil de toxicité des molécules testées.

Lors de la planification d'un essai de phase I, il peut être utile (parfois indispensable lorsqu'on utilise une approche bayésienne) d'estimer quelques courbes représentant *a priori* l'estimation de la relation entre la dose et la probabilité de toxicité *(figure 1)*. La détermination de ces graphiques est basée sur le postulat que la toxicité (et en règle générale l'efficacité) est corrélée à la dose administrée selon une fonction croissante. Les indications permettant de tracer ces courbes peuvent dériver des données précliniques, des études de pharmacocinétique/pharmacodynamie, d'essais cliniques sur des molécules de même famille thérapeutique, d'avis d'experts et de simulations réalisées à partir de toutes ces informations. Ces courbes permettent de mieux visualiser et d'optimiser la dose initiale, les paliers de dose sélectionnés et la dose correspondant *a priori* à la DMT.

Remarque

- La probabilité de TDL est classiquement représentée par une fonction de type : $p = exp^{(3+\beta x)} / (1+exp^{(3+\beta x)})$ dépendant du paramètre β.
- La *figure 1* représente quelques courbes dose/réponse répondant à ce type d'équation en fonction du paramètre β. Chaque courbe représente la probabilité de toxicité cumulée en fonction de la dose administrée. L'objectif de la méthode *Continual Reassessment Method*, décrite ultérieurement, est d'identifier le paramètre ß permettant la meilleure modélisation de la courbe réelle.

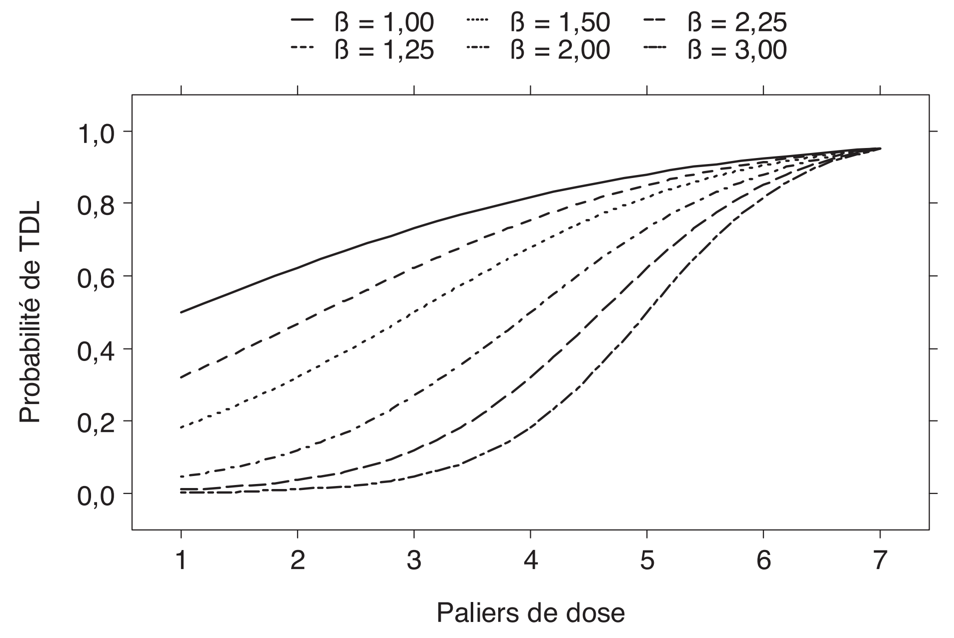


Figure 1. Exemple de courbes dose/toxicité.
TDL : toxicité dose limitante.

Choix du modèle d'escalade de dose

Le principe général d'augmentation des doses (« escalade ») est d'inclure les premiers patients au palier de dose minimal puis d'« escalader » les doses pour les patients suivants en fonction de la survenue ou non de TDL. Certaines études incluent les premiers patients à un palier de dose supérieur au palier 1 ; dans ce cas, pour les patients suivants, on pourra réaliser soit une augmentation, soit une diminution de dose en fonction des TDL observées. En règle générale, la stratégie d'escalade de dose est basée sur les toxicités observées au premier cycle administré et non pas sur l'ensemble des informations recueillies durant tout le traitement du patient.

On distingue deux types de schémas d'escalade de dose :
- les schémas basés sur des algorithmes : ce sont les plus simples et les plus utilisés ;
- les schémas basés sur un modèle de relation dose-toxicité : ils sont plus complexes et moins fréquemment utilisés malgré leur fondement statistique plus rigoureux.

Escalade de dose basée sur un algorithme

Le premier schéma basé sur un algorithme est le schéma *Up and Down*, dont les propriétés dérivent d'une chaîne de Markov de degré 1 *(figure 2)*. Ce schéma initial a ouvert la voie à de nombreuses variantes : la plus courante est le schéma *3+3* [1].

Le schéma *3+3 (figure 3)* est basé sur des considérations empiriques fondées sur le postulat que la DMT est la dose pour laquelle au moins un tiers des patients présentent une TDL. Par convention, la DR correspond à la posologie du palier précédant la DMT ; il est de règle d'y inclure au moins 6 patients. Elle produit de 0 % (0/6) à 17 % de toxicité (1/6). Simon *et al.* ont cherché à optimiser le schéma *3+3* [5]. Les principales modifications proposées étaient de ne traiter qu'un seul patient par palier et de n'augmenter le nombre de patients par cohorte qu'à l'apparition de toxicité. Les doses entre chaque palier pouvaient augmenter de 100 %, le grade de la toxicité était pris en compte dans le critère d'escalade de dose. Il était introduit la possibilité d'une escalade ou désescalade de dose intra-patient. Plusieurs simulations ont été faites : l'auteur montre qu'en utilisant ces modifications, le nombre total de patients inclus peut être réduit de 20 % à 40 %. La durée de l'étude et le nombre de patients sous-traités sont diminués. Les DMT mises en évidence semblent identiques à celles d'une escalade de dose « 3+3 » classique. En revanche, ces schémas désignés *Accelerated titration design* se montrent plus agressifs et les risques de toxicités sont plus importants.

Il existe de nombreux autres schémas d'escalade de dose basés sur un algorithme : le *Best of five* [6], le *Rolling six design* [7], le *Random walk rule* [8], le modèle de régression isotonique [9].

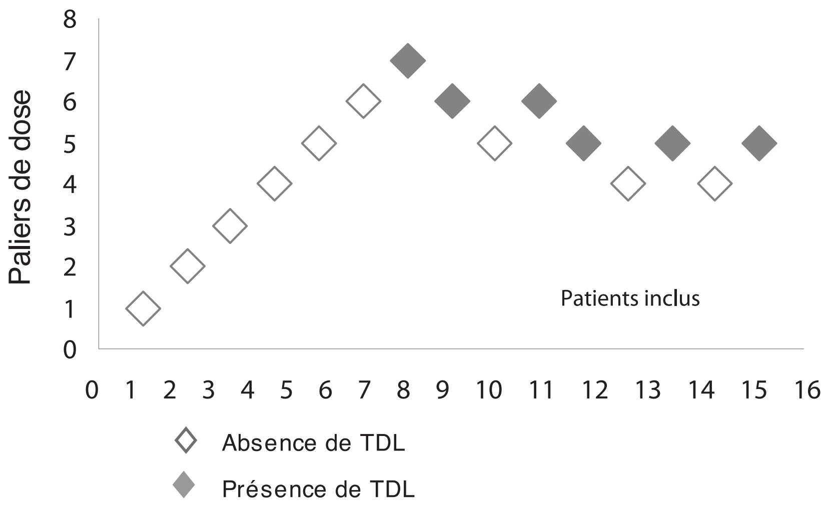


Figure 2. Schéma *Up and Down* original.
TDL : toxicité dose limitante.

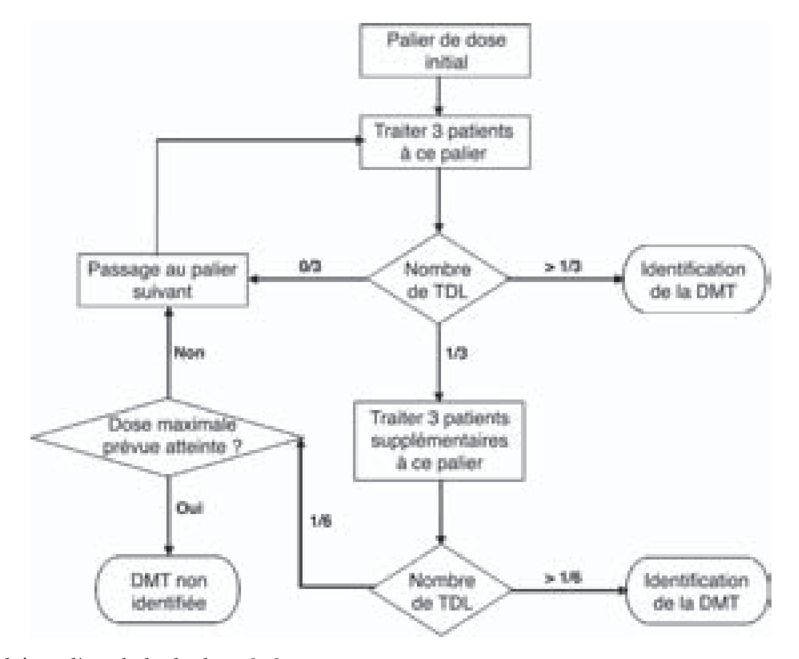


Figure 3. Schéma d'escalade de dose *3+3*.
TDL : toxicité dose limitante ; DMT : dose maximale tolérée.

Escalade de dose guidée par un modèle mathématique

Méthode de réévaluation continue

Les méthodes d'escalade de dose basées sur un algorithme sont souvent critiquées du fait qu'elles ne sont fondées sur aucune inférence statistique rigoureuse. Des études de simulation montrent que ces schémas incluent trop de patients aux paliers de dose infrathérapeutiques. Ils conduisent fréquemment à recommander une dose dont la toxicité est supérieure à la limite acceptable fixée à 33 %. De plus, l'estimation de la DMT manque de précision et l'augmentation du nombre de patients inclus ne s'accompagne pas d'une convergence de l'estimation vers la vraie valeur de la DMT.

La méthode de réévaluation continue (CRM pour *Continual Reassessment Method*) [10] est une méthode guidée par les estimations des probabilités de toxicité aux différentes doses obtenues par un modèle statistique. Il est démontré qu'au fur et à mesure que les patients s'accumulent dans l'étude, ces modèles convergent asymptotiquement vers la dose ciblée. Ils entraînent moins d'allocations de patients à des doses infrathérapeutiques et utilisent plus efficacement l'information accumulée. Le principe de base est que parmi une famille de courbes dose/toxicité relatives à un traitement *(figure 1)*, il en existe une qui correspond mieux aux données que les autres. Il convient, au préalable, de choisir une probabilité de toxicité acceptable pour la DR (par ex. 20 %, 25 % ou 30 %) et de préciser une règle d'arrêt de l'étude dès la rédaction du protocole. Il existe de nombreuses méthodes plus ou moins complexes conditionnant l'arrêt de la CRM. Chaque nouveau patient est inclus à la dose considérée comme étant la meilleure estimation de la DR en utilisant toutes les informations de tous les patients préalablement inclus. La courbe dose/toxicité est réestimée à chaque nouvel événement survenant sur les patients inclus. Deux méthodes d'estimation des paramètres du modèle ont été proposées : soit une approche bayésienne [10], soit une approche par la méthode du maximum de vraisemblance (CRML) [11]. La procédure bayésienne requiert, lors de la planification de l'étude, de spécifier une distribution *a priori* du paramètre du modèle. Cette estimation nécessite d'avoir des informations sur la molécule testée, ce qui est rarement le cas pour les molécules de première administration chez l'homme. Dans ce cas, l'opinion d'experts est une source de savoir essentielle permettant cette étape préliminaire. Paoletti et Kramar ont montré qu'en cas d'erreur sur l'estimation *a priori* du paramètre du modèle, les conséquences peuvent être sévères [12]. La méthode du maximum de vraisemblance évite cette étape délicate de la construction d'une distribution *a priori*. Sa principale limite est que le maximum de vraisemblance n'est évaluable que lorsqu'au moins une TDL et une non-TDL ont été observées. Un schéma en deux étapes est nécessaire : dans un premier temps, les doses pourront être escaladées selon un algorithme prédéfini que l'on arrêtera dès l'apparition de la première toxicité pour basculer vers une escalade reposant sur une approche CRML.

Intérêt et limites

Il apparaît que la méthode CRM traite plus de patients aux doses proches de la DR et moins de patients aux doses infrathérapeutiques que les schémas traditionnels [13]. Elle permet un ciblage rapide et une estimation précise de la DR. Cependant, elle peut entraîner des surdosages [13, 14] et ne semble pas permettre de gain de temps [13]. Au total, les schémas d'escalade de dose type CRM semblent plus efficients que ceux fondés sur un algorithme empirique. Pourtant, ils

demeurent encore très peu utilisés en pratique. Le principal obstacle est que cette méthode nécessite la collaboration étroite d'un biostatisticien et l'utilisation de logiciels spécifiques après le traitement de chaque patient.

Il existe de nombreuses autres méthodes d'escalade de dose basées sur un modèle statistique :
• l'*Extended CRM* qui est une CRM précédée d'une escalade de dose traditionnelle type « 3+3 » [15] ;
• la CRM en deux étapes incluant la possibilité de stopper l'étude très précocement dans le cas où le traitement serait très toxique [16] ;
• la CRM stratifiée permettant de diviser la population étudiée en groupes à risques distincts de toxicités [17] ;
• l'APC-DET *(Adoptive Predictive Control for Dose Escalation Trial)* tenant compte du type et du grade de la toxicité [18] ;
• le TITE-CRM *(Time To Event Continual Reassessment Method)* introduisant la notion de délai d'apparition de l'événement [19] : cette technique peut être intéressante dans le cas des traitements pour lesquels la toxicité n'est pas immédiate mais différée dans le temps. Elle apparaît relativement bien adaptée à la radiothérapie [20] ;
• l'*EWOC (Escalation With Overdose Control)* dont l'objectif est, selon une approche bayésienne, d'atteindre la DMT aussi vite que possible avec la contrainte que la proportion de patients recevant une dose supérieure à la DMT ne soit pas supérieure à une valeur seuil prédéterminée [21]. Comparé à la CRM, ce schéma réduirait le nombre de patients traités à des doses toxiques et estimerait la DMT aussi précisément que la CRM.

Escalade de dose guidée par la pharmacocinétique

Le principe de l'escalade de dose guidée par la pharmacocinétique (EDGP), décrit par Collins *et al.* dès 1985 [22], est d'atteindre le plus rapidement possible la concentration plasmatique efficace de la molécule étudiée. L'auteur montre que l'exposition plasmatique du patient à la molécule est un paramètre plus précis que la dose administrée. La mesure la plus appropriée de ce paramètre est l'évaluation de l'aire sous la courbe (ASC) de la concentration plasmatique au cours du temps. L'impact de cette technique est de permettre une escalade de dose plus rapide. Schématiquement, l'escalade de dose est rapide avant l'obtention de toxicité ou de l'ASC cible, ensuite l'escalade de dose est plus lente jusqu'à l'obtention de TDL ou de l'ASC cible. Piantadosi et Liu ont décrit en 1996 une méthode CRM dont l'escalade de dose est basée sur un critère de jugement pharmacocinétique [23] ; ce schéma n'a jamais été utilisé en pratique.

Phase I évaluant une thérapie ciblée

Le principe des essais de phase I basé sur l'attitude très caricaturale du *the more is better* est à remettre en question lorsque l'on considère les thérapies ciblées. En effet, théoriquement, le mécanisme d'action de ces molécules suggère qu'à partir d'un certain niveau de dose, on atteint un plateau ; toutes les cibles sont saturées et il n'est plus nécessaire d'augmenter les doses car

l'effet thérapeutique est maximal. En revanche, la toxicité pourrait devenir intolérable, pouvant même entraîner le rejet de molécules efficaces. De fait, dans le cas des études de phase I sur des molécules d'actions ciblées répondant à ce concept d'effet plateau, il est conseillé d'évaluer, pour chaque nouveau patient, la survenue de TDL ainsi que l'efficacité clinique et/ou pharmacologique. On choisira un schéma de type phase I-II évaluant conjointement la toxicité ainsi que l'efficacité (*cf.* section « Phase I-II ») ou bien on optera pour deux schémas d'escalade en parallèle : un schéma basé sur la toxicité et un autre basé sur l'efficacité tels que ceux développés par Hunsbergern *et al.* [24]. Le premier schéma d'Hunsbergern baptisé proportion « 4/6 » est empirique et s'inspire du schéma *3+3* classique (*cf.* Exemple 1). Le second schéma se propose de « cibler le plateau » : l'étude cesse lorsque la pente calculée de la courbe dose/réponse devient nulle (*cf.* Exemple 2).

Exemple 1 : proportion « 4/6 » et « 5/6 » [24]

On considère deux probabilités de réponse $p_0 = 0{,}30$ (taux insuffisant) et $p_1 = 0{,}80$ (taux satisfaisant) : la première entraîne la poursuite de l'escalade de dose, la seconde est cliniquement intéressante et détermine l'arrêt de l'escalade de dose.
Il existe également un autre schéma baptisé proportion « 5/6 » basé sur $p_0 = 0{,}40$ et $p_1 = 0{,}90$.
L'algorithme pour la méthode proportion « 4/6 » est le suivant :
- on escalade les doses par cohorte de 3 patients ;
- on inclut une cohorte de 3 patients supplémentaires au palier considéré dès que l'on observe ≥ 1/3 réponses ;
- la DR est celle qui entraînera ≥ 4/6 de réponses.

La probabilité de continuer l'escalade de dose est élevée lorsque le taux de réponses réel est proche de p_0, elle est faible lorsque le taux de réponses est proche de p_1.

Phase I évaluant une association de molécules

L'association de plusieurs molécules est un concept très attractif sur le plan théorique mais qui demeure fort complexe quant au choix des paliers de dose ou du modèle d'escalade de dose approprié. L'objectif est d'induire une activité thérapeutique additive, voire synergique, entre les diverses molécules en présence, sans observer une augmentation intolérable de la toxicité. Bien qu'elles ne préjugent pas du résultat de l'essai, des études précliniques préliminaires confirmant les hypothèses théoriques doivent impérativement être réalisées afin de disposer d'un rationnel solide. Divers schémas d'escalade de dose ont été proposés, soit en fixant la dose d'une molécule et en augmentant la seconde, soit en fixant alternativement la dose de l'une ou l'autre, voire en augmentant les doses des deux molécules simultanément. Les molécules testées ayant déjà été utilisées sur l'homme, la dose initiale et les paliers de doses peuvent être évalués assez précisément. En règle générale, 2 à 3 paliers suffisent pour répondre à l'objectif, une approche de type *3+3* peut raisonnablement être envisagée. Elle permet par ailleurs d'obtenir des informations sur les interactions pharmacologiques des molécules en présence [26]. Il est possible d'utiliser une escalade de dose de type CRM [27] ou un modèle d'inférence bayésienne [28].

Exemple 2 : cibler le plateau [25]

Il s'agit d'une étude de phase I menée sur un inhibiteur de la dihydropyrimidine déshydrogénase (DPD : enzyme responsable du catabolisme de la capécitabine) destiné à être associé à la capécitabine afin d'en potentialiser les effets. L'escalade de dose est basée sur un critère binaire évaluant d'une part la toxicité (schéma proche du *3+3* classique) et d'autre part l'efficacité pharmacologique (mesure de l'inhibition de l'activité de la DPD). Le critère d'arrêt pharmacologique retenu est de stopper l'escalade de dose dès que l'on obtient le même effet d'inhibition de DPD entre deux paliers successifs. Les résultats montrent une courbe effet/dose sigmoïde avec l'obtention d'un plateau théorique autour de 75 à 100 mg entraînant une inhibition de DPD de l'ordre de 90 % *(figure 4)*.

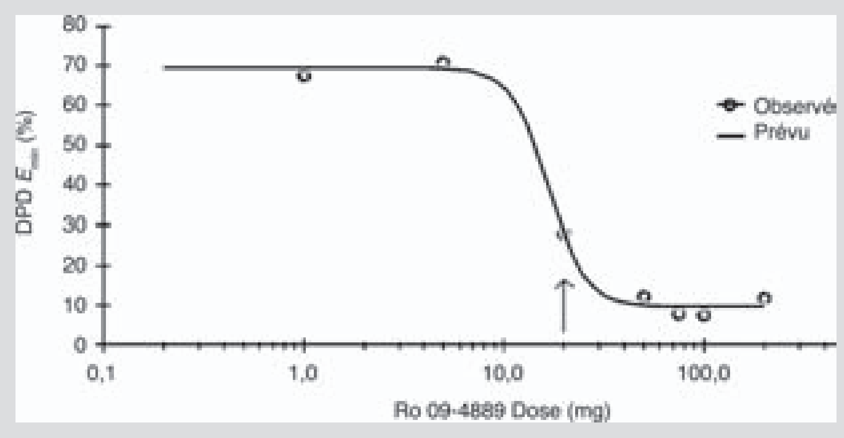


Figure 4. Exemple d'une courbe avec un effet plateau.
DPD : mesure de l'activité de la dihydropyrimidine déshydrogénase.

Phase I-II

L'objectif d'une étude de phase I-II est de réaliser simultanément une étude de phase I de toxicité et une étude de phase II d'efficacité. Ces études permettent de réduire le coût et la durée du développement de la molécule. Sur le plan éthique, les patients sont inclus dans des études dont l'objectif est l'évaluation du rapport efficacité/toxicité. De nombreux auteurs ont développé des schémas spécifiques cherchant à mettre en évidence la probabilité d'efficacité maximale sous des conditions restrictives de toxicité. Ces schémas se distinguent les uns des autres par leurs inférences statistiques sous-jacentes et également en fonction du critère de jugement de l'efficacité ou de son critère de substitution [29]. Récemment, Houédé *et al.* [30] ont développé un schéma d'essai de phase I-II d'association comprenant un agent cytotoxique et un agent de thérapie ciblée. Des simulations sous différentes hypothèses, à partir d'études cliniques, ont permis de valider cette stratégie. Il existe de nombreux logiciels permettant de faire des simulations et de conduire de telles études. Cependant, la modélisation mathématique complexe et la nécessité d'outils informatiques rebutent souvent les cliniciens à l'utilisation de ces schémas. En pratique,

il apparaît que ces schémas théoriques de phase I-II ne sont quasiment jamais appliqués. Quelques logiciels et liens Internet pour les essais de phase I sont donnés dans le chapitre VI.5 (« Les logiciels », page 370).

Conclusions

L'objectif des essais de phase I est d'estimer rapidement et précisément la dose recommandée du traitement à l'étude. Aucun schéma de phase I ne peut être considéré comme universel, différentes situations sont possibles, cette conclusion tente d'apporter une aide dans la sélection du schéma le mieux adapté au traitement considéré.

Le choix de la dose initiale et des paliers de dose est une étape primordiale, elle conditionne la qualité de l'étude. Elle est guidée par les études précliniques et cliniques, sur la molécule ou ses analogues thérapeutiques [3]. Pour les molécules déjà testées sur l'homme, le palier de dose initial ainsi que les suivants peuvent être choisis très pertinemment, le nombre de paliers testés sera généralement limité. Pour les essais sur des associations, il peut être judicieux de prioriser l'une des deux molécules. L'option de concevoir des paliers de dose « négatifs » permettant une désescalade des posologies peut être également intéressante.

Le choix du modèle d'escalade de dose peut être défini en fonction des produits testés et des résultats attendus. Dans le cas où l'on ne dispose que de très peu d'informations sur la molécule, on pourra débuter avec une escalade *3+3* [15]. Dès les premiers événements, on pourra basculer vers un modèle CRM estimé par la vraisemblance [11]. Dans le cas où les toxicités sont différées, on pourra choisir un modèle type TITE-CRM [19]. Lorsque la pathologie étudiée nécessite une stratification des patients sur certaines variables, on pourra choisir un modèle de type APC-DET [18]. Dans le cas des thérapies ciblées, il apparaît nécessaire d'introduire des critères d'arrêt basés sur l'efficacité. On pourra utiliser un schéma adaptatif de phase I-II évaluant simultanément l'efficacité et la toxicité [29]. Lorsque l'évaluation de la toxicité et de l'efficacité n'est pas concomitante, on peut envisager un critère principal d'arrêt de toxicité et un critère secondaire d'efficacité de type proportion « 4/6 » [24]. On pourra choisir un schéma ciblant un effet pharmacologique dans le cas où ce paramètre est fortement corrélé à l'efficacité du traitement [22, 25]. Dans le cas des molécules déjà testées sur l'homme pour lesquelles on dispose d'informations préalables, l'escalade de dose pourra suivre un modèle type CRM avec estimation bayésienne afin de ne perdre aucune information connue sur le traitement et de conclure le plus rapidement possible [10]. Si la dose initiale et les paliers suivants sont choisis très pertinemment un schéma *3+3* peut être efficace à condition de garantir une précision satisfaisante de la dose recommandée en incluant un nombre suffisant de patients à ce palier [26]. Lorsque l'on souhaite évaluer les propriétés pharmacocinétiques et pharmacodynamiques des molécules testées ou dans le cas des associations de drogues pouvant interagir entre elles, il sera nécessaire d'inclure plusieurs patients à chaque palier de dose afin d'avoir suffisamment de données à exploiter. Dans le cas des associations thérapeutiques, on pourra éventuellement avoir recours à des modèles plus complexes très peu utilisés en pratique [27, 28, 30].

Références

1. Storer BE. Design and analysis of phase I clinical trials. *Biometrics* 1989 ; 45 : 925-37.

2. Carter SK. Clinical trials in cancer chemotherapy. *Cancer* 1977 ; 40 : 544-57.

3. Reagan-Shaw S, Nihal M, Ahmad N. Dose translation from animal to human studies revisited. *FASEB J* 2008 ; 22 : 659-61.

4. Freireich EJ, Gehan EA, Rall DP, Schmidt LH, Skipper HE. Quantitative comparison of toxicity of anticancer agents in mouse, rat, hamster, dog, monkey, and man. *Cancer Chemother Rep* 1966 ; 50 : 219-44.

5. Simon R, Freidlin B, Rubinstein L, Arbuck SG, Collins J, Christian MC. Accelerated titration designs for phase I clinical trials in oncology. *J Natl Cancer Inst* 1997 ; 89 : 1138-47.

6. Storer BE. An evaluation of phase I clinical trial designs in the continuous dose-response setting. *Stat Med* 2001 ; 20 : 2399-408.

7. Skolnik JM, Barrett JS, Jayaraman B, Patel D, Adamson PC. Shortening the timeline of pediatric phase I trials: The rolling six design. *J Clin Oncol* 2008 ; 26 : 190-5.

8. Durham SD, Flournoy N, Rosenberger WF. A random walk rule for phase I clinical trial. *Biometrics* 1997 ; 53 : 745-60.

9. Leung DH, Wang Y. Isotonic designs for phase I trials. *Control Clin Trials* 2001 ; 22 : 126-38.

10. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase 1 clinical trials in cancer. *Biometrics* 1990 ; 46 : 33-48.

11. O'Quigley J, Shen LZ. Continual reassessment method: A likelihood approach. *Biometrics* 1996 ; 52 : 673-84.

12. Paoletti X, Kramar A. A comparison of model choices for the continual reassessment method in phase I cancer trials. *Stat Med* 2009 ; 28 : 3012-28.

13. Ahn C. An evaluation of phase I cancer clinical trial designs. *Stat Med* 1998 ; 17 : 1537-49.

14. Garrett-Mayer E. The continual reassessment method for dose-finding studies: A tutorial. *Clin Trials* 2006 ; 3 : 57-71.

15. Moller S. An extension of the continual reassessment methods using a preliminary up-and-down design in a dose finding study in cancer patients, in order to investigate a greater range of doses. *Stat Med* 1995 ; 14 : 911-22 ; discussion 23.

16. Zohar S, Chevret S. Phase I (or phase II) dose-ranging clinical trials: Proposal of a two-stage Bayesian design. *J Biopharm Stat* 2003 ; 13 : 87-101.

17. O'Quigley J, Shen LZ, Gamst A. Two-sample continual reassessment method. *J Biopharm Stat* 1999 ; 9 : 17-44.

18. Meille C, Gentet JC, Barbolosi D, Andre N, Doz F, Iliadis A. New adaptive method for phase I trials in oncology. *Clin Pharmacol Ther* 2008 ; 83 : 873-81.

19. Cheung YK, Chappell R. Sequential designs for phase I clinical trials with late-onset toxicities. *Biometrics* 2000 ; 56 : 1177-82.

20. Normolle D, Lawrence T. Designing dose-escalation trials with late-onset toxicities using the time-to-event continual reassessment method. *J Clin Oncol* 2006 ; 24 : 4426-33.

21. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat Med* 1998 ; 17 : 1103-20.

22. Collins JM, Grieshaber CK, Chabner BA. Pharmacologically guided phase I clinical trials based upon preclinical drug development. *J Natl Cancer Inst* 1990 ; 82 : 1321-6.

23. Piantadosi S, Liu G. Improved designs for dose escalation studies using pharmacokinetic measurements. *Stat Med* 1996 ; 15 : 1605-18.

24. Hunsberger S, Rubinstein LV, Dancey J, Korn EL. Dose escalation trial designs based on a molecularly targeted endpoint. *Stat Med* 2005 ; 24 : 2171-81.

25. Bellibas SE, Patel I, Chamorey E, *et al.* Single ascending dose tolerability, pharmacokinetic-pharmacodynamic study of dihydropyrimidine dehydrogenase inhibitor Ro 09-4889. *Clin Cancer Res* 2004 ; 10 : 2327-35.

26. Ferrero JM, Chamorey E, Magne N, *et al.* The raltitrexed-vinorelbine combination: A phase I pharmacokinetic and pharmacodynamic trial in advanced breast cancer. *Cancer Chemother Pharmacol* 2002 ; 50 : 459-64.

27. Kramar A, Lebecq A, Candalh E. Continual reassessment methods in phase I trials of the combination of two drugs in oncology. *Stat Med* 1999 ; 18 : 1849-64.

28. Thall PF, Millikan RE, Mueller P, Lee SJ. Dose-finding with two agents in phase I oncology trials. *Biometrics* 2003 ; 59 : 487-96.

29. Thall PF, Cook JD. Dose-finding based on efficacy-toxicity trade-offs. *Biometrics* 2004 ; 60 : 684-93.

30. Houédé N, Thall PF, Nguyen H, Paoletti X, Kramar A. Utility-based optimization of combination therapy using ordinal toxicity and efficacy in phase I/II trials. *Biometrics* 2009 ; 66 : 532-40.

Mise en œuvre d'un essai clinique de phase II

B. Pereira, A. Doussau, S. Mathoulin-Pélissier

Les essais de phase II ont pour objectif d'explorer **l'efficacité thérapeutique** de traitements évalués chez des patients. Il s'agit d'une étape charnière entre les études de phase I ayant permis de sélectionner et définir la posologie et le schéma d'administration de traitements présentant un profil de toxicité satisfaisant sur un petit nombre de patients et les études comparatives de phase III qui permettront d'évaluer l'efficacité du traitement sur beaucoup plus de patients et sur des critères à plus long terme. La conséquence directe de cette étape cruciale qu'est l'essai de phase II sera donc la poursuite ou non du développement de la stratégie thérapeutique à l'étude. Ces essais peuvent donc être vus comme une étape de filtrage aboutissant à une décision de type *go/no-go* [1]. Cette étape doit donc mettre en balance le risque d'arrêter à tort le développement de molécules ou associations pourtant bénéfiques avec le risque de poursuivre à tort dans un essai de phase III l'évaluation de traitements n'apportant pas de bénéfice (essais coûteux et entraînant potentiellement une perte de chance pour les patients participant). Ce type d'étude est parfois appelé essais *Safety and Activity* [2]. Les choix méthodologiques concernant les essais de phase II en cancérologie sont dans une phase charnière car au cours des précédentes décennies, la tendance était plutôt de minimiser le risque d'éliminer à tort des traitements efficaces car peu de molécules étaient disponibles. Actuellement, face à la quantité croissante de nouvelles molécules en cours de développement, la tendance pourrait s'inverser en faveur d'un filtre plus serré ne laissant passer en phase III que les molécules à forte probabilité de succès [1].

L'objectif principal des essais de phase II est ainsi de détecter les molécules efficaces le plus rapidement possible, en minimisant le risque de passer à côté d'un traitement actif tout en minimisant le nombre de patients traités par des molécules dont l'efficacité est insuffisante. Des implications concernant les schémas d'étude proposés devront permettre d'inclure un petit nombre de patients, d'arrêter précocement l'essai en cas d'inefficacité, de maîtriser des risques de première (alpha) et de deuxième espèce (bêta) ainsi que des critères de jugement d'efficacité évaluables à court terme (le plus souvent le taux de réponses). Il ne faut pas oublier qu'à ce stade de développement, il est encore nécessaire de recueillir les données de tolérance, voire de pharmacocinétique ; parfois, même ce critère peut être d'une importance similaire à celle de l'efficacité pour la prise de la décision quant à la poursuite ou non en phase III.

Enfin, la distinction est parfois faite entre les essais de phase IIA et IIB. Les essais de phase IIA concernent les études permettant de montrer une preuve d'efficacité *(proof of concept)*. Il s'agit le

plus souvent d'essais non randomisés à un seul bras destinés à sélectionner des traitements efficaces et éliminer ceux qui sont inefficaces. Le référentiel est alors basé sur d'autres essais ou sur l'intuition clinique et l'expérience de l'investigateur, quantifié de manière adéquate [2]. Les essais de phase IIB sont plus spécifiquement destinés à l'identification d'agents prometteurs justifiant une évaluation en phase III [3].

Schémas d'étude utilisés en phase II

Historiquement, les schémas d'études utilisés en phase II peuvent être divisés en deux catégories :
• **les essais en un seul groupe** (comparé à des données historiques) sont ceux qui ont été le plus souvent utilisés dans les essais de phase II en cancérologie mais cette approche est beaucoup plus rarement utilisée dans d'autres domaines [1]. Ces essais sont particulièrement exposés aux biais de sélection et il est donc essentiel, afin de minimiser ce biais, d'utiliser des critères d'éligibilité explicites, une taille d'étude adaptée, des critères de jugement de référence et des méthodes d'analyse rigoureuses pour évaluer les critères de jugement [2] ;
• **les essais en plusieurs groupes et randomisés** peuvent inclure le traitement évalué avec un groupe témoin ou plusieurs groupes de traitement évalués par rapport à des données historiques.

Le schéma d'étude peut reposer sur un schéma avec une taille d'étude fixe dès le départ, un schéma multi-étape – dans ce cas, l'essai peut être interrompu d'après des règles prédéfinies après chaque groupe de patients évalués dans une étape – ou un schéma séquentiel (évaluation de l'effet du traitement après chaque sujet ou paire de sujets). Comme pour toute question de recherche, différents éléments guident la stratégie dans le choix du schéma d'étude, ainsi que le choix des critères de jugement. Il est nécessaire de préciser si un traitement de référence existe et les résultats obtenus avec ce traitement (en termes d'efficacité et de toxicité), ainsi que le bénéfice clinique espéré du traitement à l'étude (en termes d'efficacité, toxicité ou les deux) [2].

Plusieurs publications ont revu les schémas d'étude utilisés dans les essais de phase II en cancérologie, avec une conclusion univoque, à savoir une très faible proportion de publications méthodologiquement satisfaisantes. En effet, moins de la moitié des essais de phase II publiés rapportent un schéma d'étude clairement identifiable [3-6].

Les schémas les plus fréquemment utilisés dans la littérature des essais des grandes revues de cancérologie sont ceux de Gehan, Fleming et Simon [6] pour les études avec un seul bras, mais le nombre d'essais de phase II randomisés croit régulièrement. De nombreuses méthodes se développent dans la littérature [7] ; nous en présenterons les principales actuellement mises en œuvre.

Les aspects méthodologiques

Choix de la population

Les patients qui participent à un essai clinique de phase II constituent un échantillon d'une population de patients susceptibles de bénéficier du nouveau traitement. Néanmoins, les critères d'inclusion sont plus restrictifs que ceux des essais de phase III, ce qui implique que ces patients ont tendance à être plus homogènes en termes de caractéristiques de la maladie. Aussi, les effets thérapeutiques observés sont souvent meilleurs et plus optimistes que les différences observées dans les essais de phase III sur une population plus large et hétérogène.

Hypothèses *a priori*, calcul de la taille d'étude, puissance

La majeure partie des approches développées pour calculer le nombre de sujets nécessaires (NSN) en phase II est fondée sur une succession d'étapes qui offrent la possibilité de prendre une décision à la fin de chacune d'entre elles : soit de poursuivre l'essai en passant à l'étape suivante, soit de stopper l'essai en concluant qu'un niveau d'efficacité minimale a été prouvé ou à une inefficacité si les résultats étaient en faveur d'un niveau d'efficacité vraiment trop faible (inefficacité maximale). L'ensemble de ces techniques dites plan multi-étapes s'avère être particulièrement intéressant. En effet, elles peuvent permettre de rejeter relativement rapidement un produit après une longue série d'échecs consécutifs tout en minimisant le nombre moyen de sujets exposés à un produit qui se verrait être inefficace.

Dans la majorité des schémas des essais de phase II, le critère de jugement principal est la réponse à un traitement définie comme un succès (critère qualitatif binaire). L'efficacité est évaluée par un paramètre π qui correspond à la probabilité réelle du succès. Considérons les données suivantes fixées *a priori* :

- k le nombre d'étapes du plan considéré ;
- p_0 la probabilité d'**inefficacité maximale** : par *inefficacité maximale*, on spécifie une probabilité de réponse p_0 qui, si elle est vraie, implique clairement que le produit ne mérite plus d'être investigué. Cette probabilité correspond habituellement aux données historiques de l'efficacité du traitement de référence que l'on cherche à supplanter ;
- p_1 la probabilité d'**efficacité minimale** : par *efficacité minimale*, on spécifie une probabilité de réponse p_1 telle que, si elle est vraie, implique que le produit a une efficacité suffisante et mérite d'être étudié.

À partir de ces définitions, on peut définir naturellement deux zones selon que le taux de succès observés π soit inférieur ou supérieur aux bornes p_0 et p_1 respectivement *(figure 1)*. En fait, ces zones d'inefficacité et d'efficacité correspondent aux zones de rejet des hypothèses nulle H_0 et alternative H_1 définies comme suit : $H_0 : \pi \leq p_0$ *versus* $H_1 : \pi \geq p_1$.

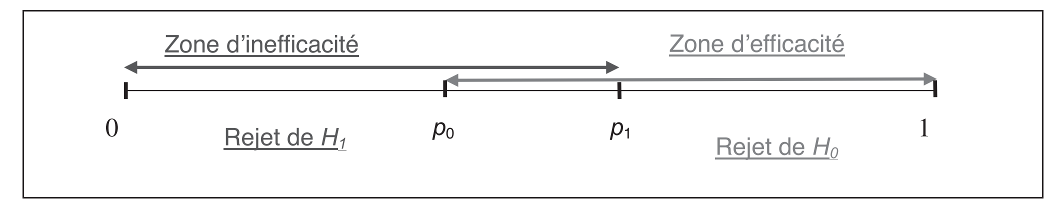


Figure 1. Définition des zones d'efficacité et d'inefficacité.

Comme pour les essais cliniques de phases I et III, le calcul du nombre de sujets d'un essai de phase II dépend également des risques de première espèce α (le plus souvent en situation unilatérale) et de deuxième espèce β qui sont fixés *a priori*. Pour obtenir des résultats cohérents et être en mesure de convaincre d'autres investigateurs de leur validité, il faut alors définir des règles de décision basées sur des résultats observés et attendus en réduisant les risques d'erreurs et les biais à un minimum.

En effet, dans ces essais, on peut voir le risque alpha comme le risque de faux positifs et le risque bêta comme le risque de faux négatifs. En phase II, il n'est presque jamais envisagé de refaire une étude négative et, inversement, une étude en faveur de l'efficacité du traitement sera dans la majorité des cas confirmée par une étude ultérieure [8]. On cherchera donc en général à minimiser le risque de faux négatif (donc en construisant un essai avec une puissance statistique élevée) en acceptant parfois un risque de faux positif un peu plus élevé. Ainsi, il n'est pas inhabituel de choisir un risque alpha de 10 % [3] puisque l'on sait que dans la plupart des cas, si l'essai de phase II est positif, l'étude aboutira à la réalisation d'un essai de phase III qui permettra de confirmer ou d'infirmer ces résultats avec un risque de première espèce plus faible (habituellement 5 %).

On peut ainsi classer les schémas d'essais de phase II selon le nombre d'étapes et le nombre de critères de jugement pris en compte :
- un critère de jugement et une étape : Fleming (1982) ;
- un critère de jugement et plusieurs étapes : plan à deux étapes de Gehan (1960) et de Simon (1989), plan de deux à cinq étapes de Fleming (1982) et plan à trois étapes de Ensign (1995) ;
- deux critères de jugement (efficacité et toxicité) et deux étapes : plan de Bryant et Day (1995).

Les plans les plus couramment utilisés seront abordés avec les principaux avantages et inconvénients *(tableau IV)*.

Particularités liées aux traitements évalués et aux critères de jugement

L'intérêt des résultats obtenus dépend de la pertinence des critères de jugement choisis. Dans le cadre classique des essais de phase II d'évaluation des traitements cytotoxiques, le taux de réponse objective est habituellement choisi, car la réponse est facilement mesurable, interprétable et pertinente cliniquement. Précocement observable, elle est mesurable de manière de plus en plus standardisée dans de nombreux cas, en particulier depuis la mise en place des critères RECIST [9].

Cependant, ce critère n'est pas toujours facile à obtenir dans les essais portant sur les traitements cytostatiques et les thérapies ciblées, les essais dans le domaine chirurgical... Dans ces situations, d'autres critères de jugements sont parfois utilisés comme l'intervalle de temps sans rechute, la survie globale ou la survie sans progression. Par ailleurs, la toxicité est systématiquement intégrée dans les critères de jugement secondaires. En effet, le nombre de patients traités préalablement en phase I étant généralement faible sur des populations sélectionnées, il est encore important, à ce stade de développement, de consolider les données de tolérance avant la phase III. Certains schémas d'étude prévoient l'intégration des critères de toxicité dans le critère de jugement principal (Bryant & Day [10]).

Contrairement aux agents cytotoxiques qui n'agissent pas uniquement sur le métabolisme des cellules tumorales mais influencent aussi celui des cellules saines, les traitements dits cytostatiques permettent de bloquer plus spécifiquement la synthèse, le fonctionnement ou la multiplication cellulaire des cellules cancéreuses. Dans ce cas, les critères de jugement et de sélection des patients établis jusqu'à présent sont remis en question pour leur pertinence. En d'autres termes, pour les thérapeutiques ciblées il faut aussi prendre en considération les éléments suivants selon le mécanisme d'action de la thérapie ciblée, le type de cancer et le stade [11] :
– l'effet positif du traitement n'implique pas forcément une réduction tumorale mais parfois simplement une stabilisation ;
– l'effet du traitement peut se manifester avec des maladies de stade peu évolué ;
– la sélection des patients est plus difficile car la cible peut être mal exprimée, voire mal connue.

Ces différents éléments expliquent la nécessité d'avoir recours à d'autres critères de jugement comme le taux de contrôle tumoral (patients répondeurs et stables) ou le temps jusqu'à progression.

Il est aussi possible de considérer des paramètres biologiques associés à la notion de biomarqueur en considérant des biomarqueurs soit en tant que critères de jugement intermédiaires (critères de substitution), soit en tant que critères paramètres du ciblage thérapeutique (critères de classification) qui sont repérés une fois que la comparaison de sujets répondeurs et non répondeurs a permis de mettre en évidence leur pouvoir prédictif de réponse au traitement (*cf.* chapitre III.2 « Facteurs pronostiques et prédictifs de la réponse à un traitement », page 149).

Les schémas avec un seul groupe

De nombreux essais cliniques utilisent encore le plan de Gehan [12] pour la planification des essais de phase II et il est important de montrer sa conception pour mieux comprendre les autres plans de phase II.

Principe du plan de Gehan

L'objectif de la première étape du plan de Gehan consiste à rejeter rapidement une molécule inefficace et arrêter l'essai si aucune réponse n'a été observée après avoir traité n_1 patients. Si, au contraire, on observe au moins une réponse, la deuxième étape avec n_2 nouveaux patients permet

d'obtenir une estimation du taux de succès avec une précision donnée. En fait, on désire limiter le risque β (fixé à l'avance, souvent 20 %) d'arrêter l'essai en concluant à l'inefficacité (souvent 5 %) si le vrai taux de réponse est supérieur ou égal à p_1. Le nombre de sujets couramment utilisé lors de cette première étape est égal à $n_1 = 14$, obtenu par la formule suivante : $n_1 = \log(\beta)/\log(1 - p_1)$. En effet, si le vrai taux de réponse est égal à 0,20, le taux d'échec est égal à 0,80. N'observer aucun succès sur 14 patients correspond à observer 14 échecs successifs. Or, si le taux d'échec pour chaque sujet est de 0,80, la probabilité d'observer 14 échecs consécutifs est égale à 0,80 x 0,80 x... x 0,80 (14 fois) soit 0,044. Si le taux de succès réel était de 0,20, on aurait relativement peu de chances (44 chances sur 1 000) de n'observer aucun succès sur 14 patients. Si on conserve l'hypothèse $p_1 = 0,20$ et on désire limiter le risque β à 1 % au lieu de 5 %, il faudra inclure 21 patients. Le risque exact est alors de 0,009. Pour $\beta = 0,05$ et $p_1 = 0,15$, il faudra inclure 19 patients dans la première étape.

L'objectif de la deuxième étape est d'obtenir une estimation du taux de réponse avec une précision donnée. Cette estimation servira de base pour les futurs essais de phase III dans le cas où la molécule serait retenue. Pour estimer le nombre de patients de la deuxième étape, des tables ont été publiées avec des paramètres suivants : l'écart-type (précision de 5 % ou 10 %), p_1 allant de 0,05 à 0,30 en pas de 0,05 et β =0,05 ou 0,10 [12]. Pour d'autres valeurs de β et p_1, le recours à des logiciels statistiques s'avère indispensable.

Le nombre de sujets à inclure dans la deuxième étape dépend également du nombre de succès observés dans la première étape. Par exemple, si on observe un seul succès parmi les premiers 14 patients (avec $\beta = 0,05$ et $p_1 = 0,20$), il faudrait inclure 46 patients supplémentaires pour estimer le taux de réponse avec une précision de ± 5 %. Ce plan, extrêmement simple, est facile à utiliser. Néanmoins, il n'est pas optimal dans le sens où la probabilité de passer à la deuxième étape reste importante malgré un produit sans grande efficacité. Par exemple, si le taux d'efficacité est seulement de 10 %, la probabilité de continuer à inclure des patients avec 14 patients est de 0,77, un chiffre beaucoup trop élevé [13].

Schéma de Fleming

Schéma de Fleming à une étape

Quels que soient le schéma proposé et le nombre d'étapes, l'objectif des plans multi-étapes reste de minimiser le risque d'éliminer une molécule active ou, *a contrario*, de rejeter rapidement des molécules insuffisamment actives et d'éviter ainsi d'y exposer inutilement des patients. Dans ce sens, commençons par nous intéresser à un schéma relativement simple : le plan de Fleming à une seule étape.

Principe et calculs

Le nombre total de sujets reposant sur une loi de Bernoulli de probabilité p et de variance p (1 – p) est obtenu comme suit :

$$N = z_{1-\alpha}\sqrt{(p_0(1-p_0))} + z_{1-\beta}\sqrt{(p_1(1-p_1))}\Big/(p_1 - p_0)^2$$

On rejettera l'hypothèse H_0 ($\pi \leq 0{,}20$) au risque $\alpha = 0{,}10$ en situation unilatérale si le nombre de succès observé (S) est supérieur ou égal à la borne supérieure R_{sup} :

$$R_{sup} = 1 + \left[Np_0 + z_{1-\alpha} \sqrt{Np_0(1-p_0)} \right]$$

Exemple

Considérons les paramètres suivants :

- le risque α est fixé à 0,10 ;
- le risque β est fixé à 0,10 et 0,20 ;
- la probabilité d'**inefficacité maximale** p_0 est égale à 0,20 ;
- la probabilité d'**efficacité minimale** p_1 est égale à 0,40.

À partir de ces quatre paramètres fixés *a priori*, on peut résumer les résultats obtenus dans le *tableau I*.

Interprétation

Dans la première partie du *tableau I* ($\beta = 0{,}10$), si on observe 10 succès ou plus parmi 33 patients, on conclut à l'efficacité de la molécule, soit un taux de réponse de 30,3 % ; sinon (9 réponses ou moins) à son inefficacité. Dans la deuxième partie du *tableau I* ($\beta = 0{,}20$), le risque de seconde espèce étant plus élevé, le nombre de sujets requis est logiquement moindre et il faut au moins 7 succès parmi 21 patients pour conclure à une efficacité suffisante de la molécule, soit 33,3 %.

En statistique, il est universellement accepté que l'on ne peut que rejeter des hypothèses (*cf.* chapitre I.3 « Compréhension des tests statistiques », page 20) et non les accepter. Dans les essais de phase II, les règles de décisions sont construites pour rejeter l'un ou l'autre hypothèse, nulle et alternative.

Contrairement à ce que l'on pense, lorsqu'on observe 6 succès ou moins sur 21 patients (taux de 28,75 % avec 6 succès), on ne peut pas dire qu'on rejette la molécule car le taux est inférieur à 20 %, alors qu'on observe un taux de 28,75 % avec 6 succès ; ce serait contraire à toute logique statistique car cela voudrait dire que l'on accepte l'hypothèse nulle. Dans ce cas, on peut dire que le taux de réponse n'est pas supérieur à 20 % – on ne rejette pas H0 – et que le taux est inférieur à 40 % (on rejette H_1 avec un risque de seconde espèce β de 20 %).

Tableau I. Exemple d'un schéma de Fleming à une étape.

Étape	Nombre de réponses (seuils limites)/nombre total de patients inclus			
	$\alpha = 0{,}10$ et $\beta = 0{,}10$		$\alpha = 0{,}10$ et $\beta = 0{,}20$	
	Rejet de H_1 $\pi \geq 0{,}40$	Rejet de H_0 $\pi \leq 0{,}20$	Rejet de H_1 $\pi \geq 0{,}40$	Rejet de H_0 $\pi \leq 0{,}20$
1	≤ 9/33	≥ 10/33	≤ 6/21	≥ 7/21

Inversement si on observe 7 succès ou plus sur 21 patients (taux de 33,3 % avec 7 succès), on peut dire que le taux de réponse est supérieur à 20 % (on rejette H_0 avec un risque de 10 %). Cela ne veut pas dire que l'on peut accepter H_1. À la fin de l'étude, on est donc en mesure soit de rejeter H_0, soit de rejeter H_1.

On peut vérifier que :
• si H_0 est vraie ($p_0 = 0{,}20$), la probabilité d'observer 7 réponses ou plus est égale à $P(R \geq 7 | n = 21, p_0 = 0{,}20) = 0{,}109$ par la loi binomiale : c'est le risque alpha (ou faux positif) ;
• de manière similaire sous H_1 : $P(R \leq 6 | n = 21, p_1 = 0{,}40) = 0{,}20$: c'est le risque bêta (ou faux négatif).

Des tables ont été établies pour une extension du schéma de Fleming basée sur la loi binomiale exacte dans les cas de petits échantillons [14].

Schéma multi-étapes de Fleming

Des plans de Fleming à plusieurs étapes (entre deux et cinq étapes) ont également été développés. Ces plans permettent d'arrêter un essai à la fin de chacune des étapes soit en concluant à l'efficacité ou à l'inefficacité, soit en continuant l'essai en passant à l'étape suivante. Les règles d'arrêt à chaque étape sont déterminées par deux conditions : limiter le risque α de déclarer intéressante une drogue d'efficacité trop faible et limiter le risque β de rejeter une drogue d'efficacité importante.

Pour commencer, il faut choisir arbitrairement le nombre maximum d'étapes et fixer les nombres de patients à inclure à chaque étape. Il faut aussi spécifier les pourcentages d'inefficacité maximale (par ex. $p_0 = 0{,}20$), d'efficacité minimale (par ex. $p_1 = 0{,}40$) et fixer les risques d'erreur α et β. Des tables sont disponibles pour des valeurs du couple (p_0 ; p_1) et $\alpha = \beta = 0{,}05$ [15].

Exemple à trois étapes : supposons que l'on veuille réaliser un plan de Fleming à trois étapes avec les mêmes paramètres $\{\alpha ; \beta ; p_0 ; p_1\} = \{0{,}10 ; 0{,}20 ; 0{,}20 ; 0{,}40\}$ et 35 patients en tout. On planifie des analyses après 15 patients dans la 1re étape et après 10 patients dans la seconde, laissant ainsi 10 patients pour la dernière étape si elle a lieu. Les zones d'arrêt calculées pour chaque étape sont présentées dans la *figure 2*.

Les risques α et β sont distribués entre les différentes étapes de manière à ce que la somme sur les différentes étapes soit inférieure aux seuils initialement fixés. Les règles d'arrêt permettant de conclure à l'efficacité sont plus strictes à la 1re étape qu'à la seconde et à la seconde qu'à la troisième. Dans cet exemple, on arrête pour efficacité :
• à la 1re étape si on a au mois 7/15 = 47 % réponses ($\alpha_1 = 0{,}018$) ;
• à la 2e étape si on a au moins 9/25 = 36 % réponses ($\alpha_2 = 0{,}052$) ;
• et à la 3e étape si on a au moins 11/35 = 31 % réponses ($\alpha_3 = 0{,}083$).

Règles de décision à l'issue de chaque étape

- **1re étape :** on arrête l'essai si l'on a observé 2 réponses ou moins. On conclut alors à une efficacité insuffisante. En revanche, si on observe 7 réponses ou plus, on conclut à l'efficacité. Enfin, dans le cas où entre 3 et 6 réponses ont été observées, on inclut 10 patients supplémentaires lors d'une seconde étape.
- **2e étape :** moins de 6 réponses observées impliquent qu'on arrête l'essai en concluant à l'inefficacité alors que 9 réponses ou plus permettent d'arrêter l'essai en concluant à l'efficacité. Pour 7 ou 8 réponses, on continue l'inclusion des patients dans une troisième étape. On peut remarquer que l'écart entre les bornes de rejet de H_0 et H_1 permettant de poursuivre l'essai devient plus petit.
- **3e étape :** après 35 patients, on conclut en rejetant la molécule si moins de 10 réponses ont été observées ou en l'acceptant si 11 réponses ou plus ont été observées. Dans tous les cas, on pourra donc conclure après évaluation des 35 patients.

On obtient, avec ces règles, $\alpha = 0{,}083$ et $\beta = 0{,}129$. De plus, si l'essai était répété de nombreuses fois, on pourrait s'attendre à inclure en moyenne 24,8 patients si le produit est inefficace et 21,6 patients si le produit est efficace.

On peut facilement vérifier que ces résultats concordent avec les intervalles de confiance (IC) correspondants à chaque étape. Par exemple, à la 1re étape, l'IC à 96,4 % ($1 - 2\alpha_1$) de 7/15 = 47 % est égal à [0,200 ; 0,748], ce qui implique une décision de rejet de H0, car la borne inférieur de l'IC est supérieure ou égale à 0,20. De manière analogue, on peut reproduire ces calculs pour chaque étape.

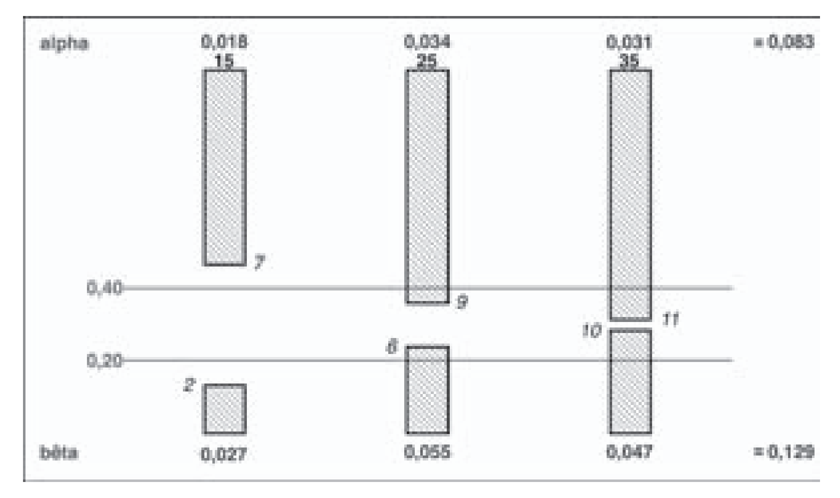


Figure 2. Exemple de design de Fleming à trois étapes et des règles de décision par étape en fonction du nombre de réponses (chiffres au milieu du graphique) ($p_0 = 0{,}20$, $p_1 = 0{,}40$, N = 35, $n_1 = 15$, n_2 et $n_3 = 10$).
Les chiffres en gras correspondent au nombre de patients inclus à chaque étape (au-dessus des blocs grisés), ceux en italiques aux seuils définissant le nombre de réponses indiquant les règles de décision. Les risques de chaque étape sont indiqués respectivement en haut pour le risque alpha et en bas pour le risque bêta ; sur la même ligne à droite est indiqué le risque cumulé sur les trois étapes.

Schéma de Simon

Simon a proposé des améliorations aux plans de Fleming en proposant un schéma à seulement deux étapes compte tenu des contraintes logistiques liées à des essais à plusieurs étapes [16]. À la différence des schémas de Fleming, le nombre de sujets dans chaque étape n'est pas fixé d'avance, **et cette approche ne prévoit pas d'arrêter l'essai pour conclure à l'efficacité à la première étape**, sauf si les règles d'arrêt de la dernière étape sont atteintes dès la première. En effet, selon Simon, il n'y a pas de problème éthique lorsqu'on continue à donner un traitement efficace. Les règles de décision à chaque étape tiennent compte des risques α et β, mais surtout ils sont déterminés d'un point de vue éthique pour minimiser le nombre moyen de sujets exposés à un produit inefficace.

Ces plans « **Optimum** » sont plus souples que les plans de Fleming et incluent en moyenne moins de sujets si le produit n'est pas suffisamment efficace. Par exemple, considérons un essai de phase II où on désire limiter ($\alpha = 0{,}10$) le risque de conclure à l'efficacité d'une molécule qui donne 20 % réponses ou moins. Par ailleurs, on souhaite limiter le risque β de conclure à son inefficacité si la molécule donne 40 % de réponses ou plus. Pour un nombre maximal de 50 sujets, le plan qui minimise le nombre moyen de sujets si le produit n'est pas efficace est présenté dans le *tableau II*.

Les règles de décision associées à cet exemple où l'on prévoit d'inclure 17 patients à la première étape et 20 patients à la seconde (37 patients au total) pour $\alpha = 0{,}10$ et $\beta = 0{,}10$ sont donc les suivantes :

- *1re étape :* on décide d'arrêter l'essai à la 1re étape si on observe entre aucune et 3 réponses (0, 1, 2 ou 3) sur 17 sujets et on conclut alors à l'inefficacité. *A contrario*, observer 4 réponses ou plus conduit à inclure 20 patients supplémentaires ;
- *2e étape :* après 37 patients, on conclut en rejetant la molécule si 10 réponses ou moins ont été observées ou en l'acceptant si 11 réponses ou plus ont été observées.

Avec ces règles, on obtient $\alpha = 0{,}0948$ et $\beta = 0{,}0967$. Si l'essai était répété, on pourrait s'attendre à ce que le nombre de sujets inclus soit de 26 en moyenne si le produit n'est pas suffisamment efficace. La probabilité d'un arrêt précoce (à la 1re étape) si le produit n'est pas efficace est de 54,9 %.

Tableau II. Exemple de schéma de Simon ($p_0 = 0{,}20$, $p_1 = 0{,}40$).

Étape	Nombre de réponses (seuils limites)/nombre de patients			
	$\alpha = 0{,}10$ et $\beta = 0{,}10$		$\alpha = 0{,}10$ et $\beta = 0{,}20$	
	Rejet de H_1 $\pi \geq 0{,}40$	Rejet de H_0 $\pi \leq 0{,}20$	Rejet de H_1 $\pi \geq 0{,}40$	Rejet de H_0 $\pi \leq 0{,}20$
1	≤ 3/17	–	≤ 6/21	–
2	≤ 10/37	≥ 11/37	≤ 7/25	≥ 8/25

D'autres plans dits « **Minimax** » ont été établis par Simon de façon à minimiser le nombre maximum de sujets exposés à un produit inefficace. Ainsi, si on reprend l'exemple précédent, le plan Minimax limite le nombre total de sujets à 33 au lieu de 35, avec des règles de décision à chaque étape correspondant à 3 réponses sur 22 patients et 7 réponses sur 33 sujets respectivement. Avec ces règles de décision, on obtient par ailleurs $\alpha = 0{,}041$, $\beta = 0{,}078$ et un nombre moyen de sujets inclus de 26,2 si le produit n'est pas suffisamment efficace.

À noter que comme ses prédécesseurs Gehan et Fleming, Simon a publié des tables pour des valeurs de p_0 allant de 0,05 à 0,30 avec un écart par rapport à p_1 de 0,15 ou de 0,20 pour des valeurs α et β égales à 0,05 et 0,10. Pour d'autres valeurs de p_0, p_1, α et β, il faut utiliser un des outils informatiques pour mettre en œuvre les plans Optimum et Minimax.

Schéma de Bryant & Day

Il peut être nécessaire de considérer simultanément deux critères de jugement pour la décision de l'intérêt d'un traitement : la toxicité en plus du critère d'efficacité. Différentes approches ont été proposées, notamment fondées sur le principe du théorème de Bayes (*cf.* chapitre I.6 « Statistiques bayésiennes », page 51). Nous nous intéressons ici au schéma de Bryant et Day [10] qui peut être vu comme approche Optimum en deux étapes de Simon intégrant conjointement comme critère de jugement principal la toxicité T et l'efficacité R. On comprend aisément que les paramètres fixés initialement ne seront pas exactement les mêmes que dans les exemples précédents. Ainsi, lorsqu'on met en place cette technique, on doit fixer non plus deux mais trois risques d'erreur :

- le risque α_R associé à l'efficacité R ;
- le risque α_T associé à la toxicité T ;
- le risque β pour R et T.

De même, un nouveau critère de jugement étant impliqué dans cette approche, il faut définir les valeurs initiales p_{T0} et p_{T1} associées aux taux de non-toxicité minimale et maximale acceptables, comme nous l'avons fait pour les deux précédents exemples avec l'efficacité p_{R0} et p_{R1}. Les conclusions fournies par le design de Bryant et Day prennent également en compte cette double caractérisation du critère de jugement et permettent de déterminer un nombre de patients associé à chacun des deux critères R et T. Ainsi, on rejettera le traitement à la fin de l'étape k ($k = 1$ ou 2) si :

- le nombre de réponses est inférieur ou égal à un seuil ;
- le nombre de patients ne présentant pas de toxicité est inférieur ou égal à un seuil.

Cette méthode a été développée en considérant une hypothèse d'indépendance entre toxicité et efficacité.

Considérons l'exemple suivant où l'on souhaite réaliser un essai de phase II en employant le schéma de Bryant et Day en désirant limiter, d'une part ($\alpha_R = 0{,}10$) le risque de conclure à l'efficacité d'une molécule qui donne $p_{R0} = 20$ % de réponses ou moins et, d'autre part, limiter le risque ($\beta = 0{,}10$) de conclure à son inefficacité si la molécule donne $p_{R1} = 40$ % de réponses ou plus. Pour ce qui est des paramètres associés à la toxicité, on fixe de manière analogue α_T à 0,10 et :

- p_{T0} = 0,60 le taux de non-toxicité inacceptable (soit 40 % de toxicité) ;
- p_{T1} = 0,80 le taux de non-toxicité acceptable (soit 20 % de toxicité).

Les résultats obtenus, avec ces paramètres, sont résumés dans le *tableau III*.

Si on définit les designs de Bryant et Day avec les paramètres $\{n_k, r_k, t_k\}$ pour les deux étapes k = 1,2, ce design peut être représenté par {24, 5, 15} pour la 1re étape et {54, 14, 36} pour la seconde.

Tableau III. Exemple de design de Bryant et Day à deux étapes.

Étape	Nombre de patients n_k	Nombre de réponses r_k	Nombre sans toxicité t_k
1	24	≤ 5	≤ 15
2	54	≤ 14	≤ 36

Pour cet exemple, avec les hypothèses fixées, il faut inclure un total de 24 patients à la 1re étape. On conclura à :
- l'arrêt pour efficacité insuffisante si on observe 5 réponses ou moins ;
- l'arrêt pour toxicité excessive si on observe 15 patients sans toxicité ou moins (soit 9 patients avec toxicité ou plus) ;
- l'arrêt pour efficacité insuffisante et toxicité excessive si on observe 5 réponses ou moins et 15 patients sans toxicité ou moins.

Dans la situation restante (6 réponses ou plus et 16 patients sans toxicité ou plus), on reprendra les inclusions avec 30 patients supplémentaires afin d'atteindre un total de 54 patients inclus. On conclura alors :
- au rejet pour efficacité insuffisante si on observe 14 réponses ou moins ;
- au rejet pour toxicité excessive si on observe 36 patients sans toxicité ou moins (soit 18 patients avec toxicité ou plus) ;
- au rejet pour efficacité insuffisante et toxicité excessive si on observe 14 réponses ou moins et 36 patients sans toxicité ou moins ;
- à la recommandation de poursuivre l'évaluation du traitement si on observe 15 réponses ou plus et 37 patients sans toxicité ou plus.

Les essais de phase II randomisés

En cancérologie, les essais de phase II non randomisés sont souvent utilisés et ces schémas sont en fait acceptables pour les traitements cytotoxiques évalués sur le critère de taux de réponses objectives, car la réduction spontanée du volume tumoral est improbable. Ces essais peuvent poser des problèmes d'interprétation liés à une comparaison implicite historique qui n'est pas toujours fiable ou à une sélection de la population (sur ou sous-estimation de l'effet). La littérature internationale insiste aujourd'hui sur la nécessité de s'orienter, si cela est faisable, vers des essais de phase II randomisés, notamment pour l'évaluation des thérapies ciblées [17, 18]. Ce type

d'essai permet d'évaluer, dans un groupe sélectionné de la même façon que le groupe témoin, le critère de jugement au même temps de l'étude. L'objectif est donc de limiter les biais de sélection liés à la population comme les biais liés à des comparaisons historiques non valides. À côté des essais de phases II randomisés avec un bras de référence, d'autres schémas d'étude sont envisagés selon la stratégie : choisir le traitement le plus efficace entre plusieurs combinaisons (doses par ex.) pour poursuivre en phase III ou choisir la population la plus pertinente pour évaluer l'effet de l'agent testé. Ce type d'essai ne permet pas de manière formelle de comparer les traitements car la puissance statistique est très faible.

Les schémas d'étude des essais de phase II randomisés

Phase II randomisée avec un groupe témoin

Dans un essai de phase II randomisé, le traitement ou l'intervention sera attribué par tirage au sort. Le taux de réponse (ou un autre critère principal) du groupe témoin renseignera sur l'existence ou non d'un biais de sélection notable. L'objectif d'un essai de phase II n'est donc pas de comparer directement les traitements ou interventions car ce type d'analyse augmente alors le taux de faux positifs [19]. En effet, ces essais sont conçus pour évaluer le traitement expérimental par comparaison avec un bras historique. Les règles de décision sont applicables dans les deux groupes. Si le taux de réponse dans le groupe témoin est proche de celui attendu (taux utilisé dans l'hypothèse nulle), on peut conclure qu'il n'existe pas de biais de sélection des patients et cela renforce donc la valeur des conclusions sur le bras expérimental. *A contrario*, si ce taux de réponse dans le groupe témoin est très différent de celui utilisé dans l'hypothèse initiale, la conclusion est moins solide et il faudrait refaire un essai. La décision pour continuer ou non en phase III dépendra des résultats observés dans les deux bras [20].

Phase II randomisée de sélection (*selection design* ou *screening design* ou « *Pick the winner* »)

Typiquement, ce schéma d'étude est dédié pour randomiser les patients entre plusieurs doses d'agents ou schéma d'administration sans bras de référence [21]. Chaque « bras » sera testé pour son activité en utilisant les critères standards des études à un bras avec une règle de sélection pour choisir le meilleur bras pour poursuivre le développement en phase III. Ce type de schéma a l'avantage de limiter les biais de sélection des patients dus, par exemple, à une amélioration des résultats possibles dans les essais séquentiels. Mais la limite dans ce type de schéma est liée au nombre de bras : plus le nombre de bras augmente, plus la probabilité de sélectionner le meilleur bras diminue ; autrement dit c'est la puissance statistique qui diminue. Les « bras » ne sont pas comparés entre eux (et ne doivent pas être comparés !). Le risque de 1re espèce est élevé : on cherche plutôt à limiter le risque de faux négatifs (risque de seconde espèce) dans ce contexte et non le taux de faux positifs (risque de 1re espèce) ; ce type d'essai n'est donc jamais un substitut à une phase III [3].

Phase II randomisée d'enrichissement

Ce schéma d'une phase II randomisée d'enrichissement *(Randomized discontinuation design)* répond à la nécessité de déterminer l'activité des agents cytostatiques qui peuvent avoir une activité limitée si on mesure le taux de réponse et pour lesquels le temps jusqu'à progression (TTP pour *Time to progression*) est un critère plus adapté [22, 23]. Dans ce type d'essai, tous les patients reçoivent le traitement évalué pendant un temps spécifié (par ex. 2 à 4 mois). Les patients avec une progression, une toxicité ou non observants durant cette période sont alors exclus de l'essai ; les patients avec une réponse continuent le traitement ; les patients restants sont randomisés entre continuer le traitement ou non, dans un essai en aveugle avec placebo (ou double placebo selon les formes galéniques disponibles) *(figure 3)*. Le critère de jugement est le taux de patients restés stables durant la période de randomisation. Les avantages de ce type d'essai sont la meilleure adhésion au principe de randomisation puisque ne sont sélectionnés que les patients qui ont un potentiel bénéfice : cette meilleure adhésion permet ainsi un recrutement plus rapide des patients. Ce schéma d'étude est plutôt adapté aux tumeurs d'évolution lente – si l'évolution est rapide, peu de patients pourront être en pratique randomisés – et un des désavantages est le possible développement d'une résistance pendant la période initiale.

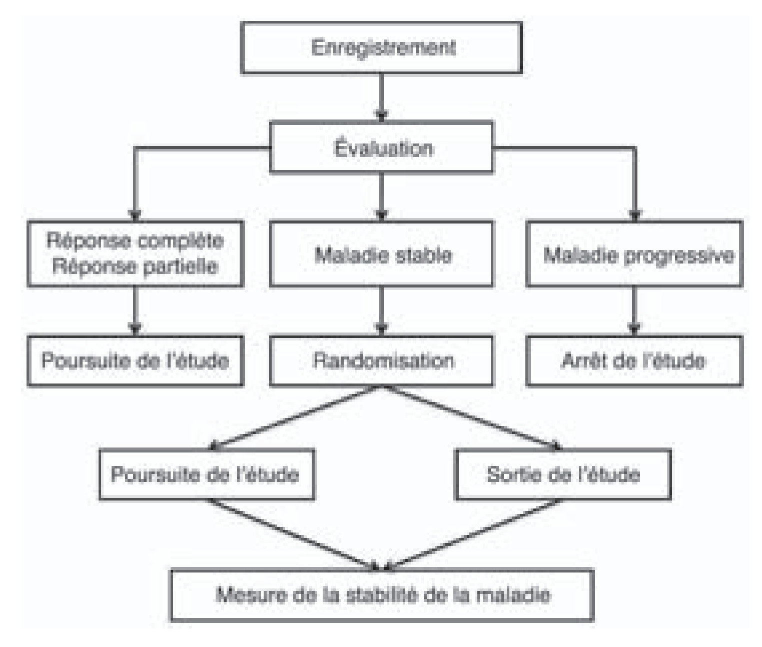


Figure 3. Essai de phase II randomisé par « enrichissement » : inclusion et suivi d'un patient.

Limites et stratégie de décision des essais de phase II randomisés

La randomisation ne peut éviter le risque lié au manque de puissance d'un essai comme cela arrive si le nombre de sujets est trop faible. Ce type d'essai peut nécessiter plus de patients à inclure et donc être un frein à son utilisation, notamment dans le contexte de cancers rares ou de populations très sélectionnées. Enfin, la randomisation dans cette étape de l'évaluation d'une molécule ou d'une stratégie peut être difficile à mettre en place par manque d'ambivalence et être donc un frein à l'inclusion : préférence du clinicien ou du patient pour un bras augmentant le taux de non-participation. Cette difficulté doit être discutée et acceptée dès l'étape de réflexion sur le plan de l'étude.

Discussion

Pour terminer, nous voudrions évoquer des discussions persistantes sur les essais de phase II en cancérologie.

Pour certains investigateurs, le passage d'emblée des phases I à III est envisageable [24]. *A contrario,* dans certaines circonstances exceptionnelles, l'autorisation de mise sur le marché peut être accordée sans un essai de phase III : maladie rare, absence de tout autre traitement, avancée clinique majeure, par exemple.

Pour d'autres investigateurs, le besoin de mettre en place plus d'essais de phase II comparatifs est une évidence car l'objectif est de pouvoir choisir les bons traitements pour poursuivre le développement. Néanmoins, l'interprétation des résultats de ces essais est fondée sur plusieurs critères d'évaluation, parfois sur un nombre limité de patients, avec souvent de nombreux traitements à évaluer et de nombreux paramètres à prendre en considération avec le choix du meilleur bras de traitement [25]. Ce point est fortement débattu et non consensuel.

En effet, il ne faut pas oublier un problème potentiel majeur d'interprétation des essais avec un seul groupe. L'objectif des essais de phase II étant de sélectionner des traitements pour lesquels poursuivre l'évaluation en phase III, la sélection implique de fait une comparaison et si la comparaison doit se faire à travers les résultats d'autres essais, alors il existe un risque important de confusion entre l'effet du traitement et l'effet de l'essai. Dans un essai avec un seul groupe, la mise en évidence de l'effet d'un traitement par rapport aux données d'essais historiques peut être liée à un réel bénéfice du nouveau traitement ou à la sélection de patients différents par rapport aux essais historiques *(treatment-trial confusing)* [26] ; de même, l'effet d'un traitement peut être neutralisé par un effet de l'essai [25]. Pourtant, pour certaines maladies ou situations rares, c'est peut-être la seule solution au début du développement d'une nouvelle molécule.

Tableau IV. Les principaux plans dans les essais de phase II [7, 19, 24].

Plan ou nom/auteur	Type	Principes	Principaux avantages	Principaux inconvénients
Fleming, 1982 [15]	Une étape	Nombre de sujets fixé à l'avance Comparaison réponse théorique/observée		Biais de sélection Contrainte d'avoir le nombre de sujets nécessaires exact pour pouvoir conclure
Gehan, 1960 [12]	Deux étapes	1re étape : arrêt si pas de succès sur les x premiers patients 2e étape : poursuite de l'essai pour obtenir le taux de réponse avec une certaine précision	Arrêt précoce si forte probabilité d'inefficacité	Biais de sélection Ne contrôle que le risque de faux négatifs Fournit précision, estimation et efficacité, ne permet pas de dire si le traitement est efficace Contrainte d'avoir le nombre de sujets nécessaires exact Contrainte de la phase d'interruption pour analyse entre les deux étapes
Simon, 1989 [16]	Deux étapes	Seuil d'inefficacité maximale tolérée pour continuer le développement et seuil d'efficacité minimale pour juger le traitement efficace *Minimax* : taille de l'échantillon d'étude la plus petite possible / *Optimum* : taille de l'échantillon de la 1re étape plus petite	Contrôle alpha et bêta	Biais de sélection Contrainte d'avoir le NSN **exact** pour pouvoir conclure, Contrainte de la phase d'interruption pour analyse entre les deux étapes

Tableau IV. (suite).				
Plan ou nom/auteur	**Type**	**Principes**	**Principaux avantages**	**Principaux inconvénients**
Fleming, 1982 [15]	Multi-étapes	K étapes ($k \leq 6$) : à chaque étape, un seuil permet d'arrêter l'essai en concluant à l'efficacité ; un seuil permet d'arrêter en concluant à l'inefficacité ; sinon on poursuit	Contrôle alpha et bêta	Biais de sélection
Bryant & Day, 1995 [10]	Deux étapes, deux critères d'évaluation	Évaluation simultanée de la toxicité et de l'efficacité Seuil d'efficacité minimum et d'inefficacité, seuil de tolérance acceptable et inacceptable, risque alpha pour l'efficacité et la tolérance à fixer	Contrôle alpha et bêta Évaluation conjointe de la toxicité et l'efficacité La méthode de minimisation basée sur l'hypothèse d'indépendance toxicité-efficacité est robuste à un écart à cette hypothèse	Biais de sélection Difficulté de définir la toxicité de manière binaire Vérification de l'hypothèse d'indépendance entre les deux critères
Test triangulaire groupé, 1990	Séquentielle	Méthode séquentielle groupée par groupe de n patients (analyse intermédiaire après obtention de chacun des n résultats), règles de décision d'arrêt ou de poursuite définies *a priori*		Besoin des n résultats avant de continuer donc d'un critère de jugement rapide d'obtention

Tableau IV. (fin).				
Plan ou nom/auteur	**Type**	**Principes**	**Principaux avantages**	**Principaux inconvénients**
Cross-over		Administration à chaque patient de l'un puis de l'autre des traitements (ordre déterminé par tirage au sort)	Moins de sujets (tumeurs rares) Étude des résistances croisées	Problèmes liés aux résistances, efficacité différente en 1re ou 2e ligne Interaction
Bayésien	Multi-étapes (*cf.* chapitre I.6, page 51)	Introduction d'une loi d'efficacité *a priori* du traitement évalué Les inclusions s'arrêtent lorsque la probabilité *a posteriori* que le traitement soit efficace par rapport au traitement historique est suffisamment petite	Durée plus courte des essais de phase II Utilisation de toutes les données pertinentes disponibles	Influence des hypothèses *a priori* sur les conclusions
Randomisé	Non comparatif	Traitement évalué et bras témoin ou deux (ou plusieurs) traitements évalués Non comparatif	Contrôle du biais de sélection Population témoin pour interprétation par rapport aux hypothèses initiales basées sur données historiques Possibilité de basculer en phase III si bras de référence conforme à ce qui était attendu	Nombre de patients plus élevés à inclure (groupe témoin)
Randomisé	Comparatif			Nécessite un compromis par rapport au risque de faux positifs Nombre de patients plus élevés à inclure (groupe témoin)

NSN : nombre de sujets nécessaires.

Références

1. Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: The importance of randomisation. *Eur J Cancer* 2009 ; 45 (2) : 275-80.

2. Piantadosi S. Sample size and power. In: Piantadosi S (ed). *Clinical Trials: A methodologic perspective*, 2nd ed. New York: John Wiley & Sons, 2005 : 251-308.

3. Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: Current status and future directions. *J Clin Oncol* 2005 ; 23 (19) : 4450-7.

4. Mariani L, Marubini E. Content and quality of currently published phase II cancer trials. *J Clin Oncol* 2000 ; 18 (2) : 429-36.

5. Perrone F, Di Maio M, De Maio E, *et al*. Statistical design in phase II clinical trials and its application in breast cancer. *Lancet Oncol* 2003 ; 4 (5) : 305-11.

6. Thezenas S, Duffour J, Culine S, Kramar A. Five-year change in statistical designs of phase II trials published in leading cancer journals. *Eur J Cancer* 2004 ; 40 (8) : 1244-9.

7. Green S. Overview of phase II clinical trials. In: Crowley J, Ankerst DP (eds). *Handbook of statistics in clinical oncology*. Boca Raton: CRC Press ; Taylor and Francis Group, 2006 : 119-29.

8. Carter S. Clinical aspects in the design and conduct of phase II trials. In: Buyse M, Staquet M, Sylvester R (eds). *Cancer clinical trials, methods and practice*. Oxford : Oxford Medical publication, 1984 : 237.

9. Eisenhauer EA, Therasse P, Bogaerts J, *et al*. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* 2009 ; 45 (2) : 228-47.

10. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995 ; 51 (4) : 1372-83.

11. Penel N, Saleron J, Lansiaux A, *et al*. Particularités méthodologiques des études cliniques appliquées à l'évaluation des thérapeutiques ciblées. *Bull Cancer* 2008 ; 95 (2) : 185-90.

12. Gehan E. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chron Dis* 1960 ; 13 (4) : 346-53.

13. Kramar A, Potvin D, Hill C. Plans expérimentaux pour l'inclusion de patients dans les essais de phase II. *Rev Epidemiol Sante Publique* 1996 ; 44 (4) : 364-71.

14. A'Hern RP. Sample size tables for exact single-stage phase II designs. *Stat Med* 2001 ; 20 (6) : 859-66.

15. Fleming TR. One sample multiple testing procedures for phase II clinical trials. *Biometrics* 1982 ; 38 : 143-51.

16. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989 ; 10 (1) : 1-10.

17. Mandrekar SJ, Sargent DJ. Randomized phase II trials : Time for a new era in clinical trial design. *J Thorac Oncol* 2010 ; 5 (7) : 932-4.

18. Seymour L, Ivy SP, Sargent D, *et al*. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the national cancer institute investigational drug steering committee. *Clin Cancer Res* 2010 ; 16 (6) : 1764-9.

19. Piedbois P. Les essais de phase II randomisés en cancérologie. *Bull Cancer* 2007 ; 94 (11) : 953-6.

20. Buyse M. Randomized designs for early trials of new cancer treatments – an overview. *Drug Information Journal* 2000 ; 34 : 387-96.

21. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep* 1985 ; 69 (12) : 1375-81.

22. Freidlin B, Simon R. Evaluation of randomized discontinuation design. *J Clin Oncol* 2005 ; 23 (22) : 5094-8.

23. Rosner GL, Stadler W, Ratain MJ. Randomized discontinuation design: Application to cytostatic antineoplastic agents. *J Clin Oncol* 2002 ; 20 (22) : 4478-84.

24. Medioni JR, Rycke YD, Asselain B. Évaluation de traitements anticancéreux par des essais de phase II : nouvelles orientations. *Bull Cancer* 2000 ; 87 (7-8) : 551-6.

25. Kramar A, Potvin D, Hill C. Multistage designs for phase II clinical trials: Statistical issues in cancer research. *Br J Cancer* 1996 ; 74 (8) : 1317-20.

26. Estey EH, Thall PF. New designs for phase 2 clinical trials. *Blood* 2003 ; 102 (2) : 442-8.

Mise en œuvre d'un essai clinique de phase III

S. Mathoulin-Pélissier, A. Kramar

Au cours de ces 25 années, les bases de l'expérimentation scientifique en médecine ont été posées et l'essai thérapeutique randomisé a été affirmé comme un dogme incontournable. Il s'agit d'études cliniques dont l'objectif est la comparaison de plusieurs (au minimum deux) approches (ou stratégies) thérapeutiques. Le traitement est attribué par tirage au sort.

Quelques dates

Le concept du tirage au sort a été introduit par RA. Fisher en 1923 [1] et le premier essai thérapeutique contrôlé a été proposé en 1931 [2]. Le terme « placebo » a été employé par HS. Diehl en 1938 [3], et la stratégie du double aveugle a été utilisée par le *Medical Research Council* en 1948 [4]. Le premier essai randomisé en cancérologie a été mené en 1954 par l'Institut national du cancer américain chez des patients atteints de leucémies.

Un essai thérapeutique de phase III doit apporter la preuve de l'efficacité du traitement testé et/ou de sa supériorité par rapport à ceux existants. L'essai de phase III doit démontrer soit l'efficacité d'un nouveau traitement par rapport à un placebo (substance dénuée d'activité pharmacologique mais perçue par le patient comme un médicament), soit (le plus souvent) sa supériorité par rapport à un traitement de référence (traitement considéré comme le meilleur ou reconnu par l'usage). L'essai de phase III doit permettre l'imputation causale : « c'est bien le nouveau traitement qui a entraîné l'amélioration avec un minimum d'erreurs sur la conclusion de l'essai, que ce soit le risque de faux positif (α) ou le risque de faux négatif (β) ».

L'essai thérapeutique randomisé (ou contrôlé) est la seule méthode reconnue, capable de démontrer la supériorité, si elle existe, d'un traitement par comparaison à un autre, puisque l'attribution du traitement est faite de manière aléatoire. Les deux groupes de sujets recevant les deux traitements à l'étude sont alors comparables pour toutes leurs caractéristiques connues, mesurables ou non, ou inconnues.

Rappel : les grands principes de l'essai thérapeutique de phase III

- Prospectif
- Comparatif : c'est-à-dire *versus* traitement de référence ou placebo
- Randomisé : comparabilité initiale
- En double aveugle
- Sans données manquantes et l'analyse principale en intention de traiter

Les essais de phase III sont ainsi considérés comme l'étape capitale pour l'obtention d'une autorisation de mise sur le marché (AMM) pour les traitements médicamenteux. Le terme d'essai pivot *(pivotal study)* est alors utilisé et la présence de cet essai est indispensable dans le dossier d'enregistrement d'AMM du médicament.

Nous renvoyons le lecteur qui souhaite approfondir la compréhension des essais thérapeutiques vers des ouvrages ou publications de référence [5-7]. Par ailleurs, ce chapitre est complété par des chapitres complémentaires : VI.2 « Modalités de randomisation » (*cf.* page 344) et VI.3 « Les comités indépendants de surveillance des essais thérapeutiques : rôle et modalités de fonctionnement » (*cf.* page 354).

Les schémas d'étude

Selon l'objectif

Les objectifs d'un essai de phase III en cancérologie peuvent être :
– de déterminer l'efficacité d'un nouveau traitement comparé au traitement standard (plus rarement à un placebo) : essai de supériorité ;
– ou de déterminer si un nouveau traitement est aussi efficace que le traitement standard, mais associé avec moins de toxicité ou avec une meilleure qualité de vie (essai d'équivalence ou de non-infériorité).

Essai de supériorité

Le principe d'un essai de supériorité est de tester l'hypothèse que le traitement expérimental est supérieur au traitement de référence (ou placebo parfois en cancérologie) sur la base d'un objectif cliniquement pertinent (par ex. avec un critère de survie).

Ainsi, l'essai de supériorité nécessite la définition d'une différence minimale d'efficacité que l'on veut mettre en évidence avec une puissance suffisante (calcul de taille d'échantillon).

Essai de non-infériorité

Les essais de non-infériorité *(non-inferiority trial)*, parfois appelés par abus de langage « essais d'équivalence » *(equivalence trial)*, deviennent de plus en plus fréquents dans l'évaluation clinique des nouveaux traitements.

Ce type d'essai fait appel à une méthodologie et à des techniques statistiques dont le développement est relativement récent et peu connu. De ce fait, des nouveaux traitements peuvent être acceptés sur la base d'essais d'équivalence discutables par méconnaissance des pièges et des spécificités de ce type d'étude. En particulier, le processus décisionnel qui leur est attaché nécessite l'introduction d'**un seuil d'équivalence choisi avec beaucoup de discussion**, car l'interprétation des résultats de l'essai repose grandement sur la valeur de ce seuil [8].

Les conclusions de ces essais sont aussi très souvent mal interprétées. Malgré les apparences, ce type d'essai ne permet pas de conclure que le traitement étudié a une efficacité identique à celle du traitement de référence mais simplement qu'il a une efficacité suffisante. Comme nous le verrons par la suite, les méthodes disponibles permettent seulement de raisonnablement éliminer la possibilité que le traitement étudié soit nettement moins efficace que le traitement de référence. Ces techniques permettent d'exclure que le nouveau traitement entraîne une perte d'efficacité supérieure à une certaine limite, fixée *a priori*. Ainsi, à l'issu d'un essai de non-infériorité concluant, la seule chose qui soit acquise (avec un risque alpha d'erreur de 5 %) est que cette perte d'efficacité est inférieure à la limite que les investigateurs sont prêts à perdre compte tenu des avantages qu'offre le nouveau traitement par ailleurs.

Selon le plan ou l'unité d'analyse

Essai en parallèle

L'essai en deux groupes parallèles, appelé aussi en bras parallèles *(parallel groups* ou *parallel arms)*, est l'archétype de l'essai thérapeutique.

Principe

Le traitement étudié est comparé à un traitement contrôle (placebo ou traitement actif) à l'aide de deux groupes de patients constitués par randomisation de façon contemporaine et suivis **en parallèle**.

Explication

La majorité des essais cliniques randomisés utilisent une approche parallèle où l'on évalue l'efficacité de deux ou plusieurs traitements. Pour ces études, chaque participant est affecté par randomisation à un bras thérapeutique à deux choix ou plus :
• deux ou plusieurs traitements concurrents sont comparés à un même contrôle (placebo ou traitement actif) et/ou entre eux ;
• un traitement étudié est comparé à plusieurs contrôles, par exemple son placebo et un traitement actif.

Essai croisé ou chassé-croisé

Principe

Dans un essai croisé *(cross-over design)*, chaque patient reçoit tous les traitements de l'essai, administrés lors de périodes successives.

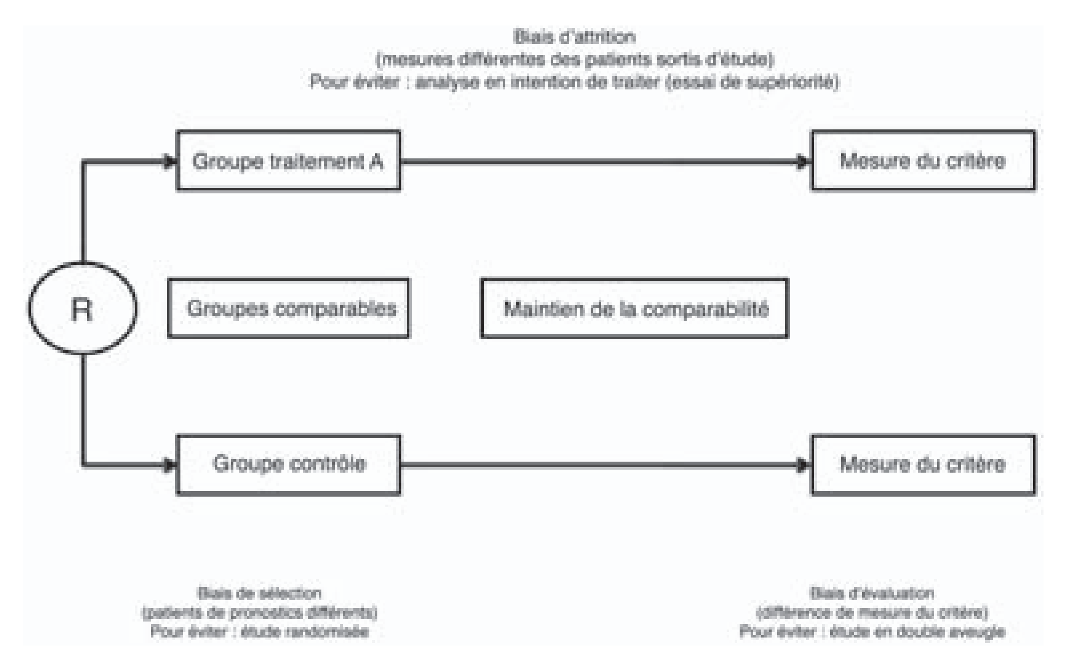


Figure 1. Essai clinique en parallèle : la randomisation, l'insu et l'analyse en intention de traiter pour limiter les biais.

Explication

Dans ce type d'essai comparant deux traitements A et B, tous les patients reçoivent de façon aléatoire soit la séquence A puis B (G1), soit la séquence B puis A (G2) *(tableau I)*. On compare donc, pour chaque patient, l'effet du traitement A lorsque le patient était sous le traitement A à l'effet du traitement B lorsqu'il était sous le traitement B. Cette approche permet de minimiser les variations propres à chaque individu puisque chaque participant est son propre contrôle. Ce schéma permet donc de diminuer la taille d'échantillon requise pour une même différence recherchée et les mêmes risques d'erreur. Cependant, on peut difficilement utiliser cette approche pour les traitements curatifs car il ne faut pas que le traitement de la première période ait encore un effet pendant le second traitement ; de même, ce schéma n'est pas optimal si la maladie n'est pas stable au cours du temps car il est nécessaire que le traitement soit utilisé dans le même contexte de sévérité de la maladie.

La valeur du critère de jugement est mesurée à la fin de chaque période. Chaque patient produit donc une mesure du critère de jugement sous le traitement étudié et une avec le traitement contrôle. Ce schéma est plus facilement adapté à l'évaluation des effets chroniques, comme des migraines, des troubles digestifs, ou à l'évaluation des constantes biologiques.

Tableau I. Exemple d'un plan en *cross-over*.		
	Groupe	
Période	**Groupe 1 (G1)**	**Groupe 2 (G2)**
Première période (P1)	Traitement A	Traitement B
Deuxième période (P2)	Traitement B	Traitement A

Plan ou essai factoriel

Principe

Le plan factoriel *(factorial design)* permet d'effectuer simultanément des comparaisons de plusieurs traitements pris seuls ou en association. Dans un plan 2x2, les patients sont randomisés en quatre groupes, chacun recevant ou non les traitements A et B. la première comparaison est celle du traitement A à son contrôle (ou placebo), la seconde celle du traitement B à son contrôle (ou placebo).

Explication

Les patients d'un essai factoriel seront randomisés une première fois entre A et son contrôle, puis une seconde fois, sans tenir compte de la nature du premier traitement reçu, entre B et son contrôle. Ces deux randomisations simultanées créent en fait quatre groupes de patients : 1/4 des patients recevront le traitement A et le traitement B (A+B+), 1/4 recevront le traitement A et le contrôle du traitement B (A+B-) ; 1/4 recevront le contrôle du traitement A et le traitement B (A-B+) ; 1/4 recevront le contrôle de A et le contrôle de B (A-B-) *(tableau II)*.

En pratique, la randomisation des patients se fait directement entre ces quatre groupes afin d'assurer l'équilibre des effectifs. Au total, la moitié des patients reçoit le traitement A et l'autre moitié le contrôle de A. De même, la moitié des patients reçoit le traitement B et l'autre moitié le contrôle de B.

Le nombre de groupes augmente de façon exponentielle pour chaque nouveau traitement étudié ; ainsi, un essai évaluant trois traitements devra avoir 2x2x2 = 8 groupes différents.

L'avantage principal des essais qui utilisent un plan factoriel est que l'on peut mener deux ou plusieurs études simultanément avec moins de patients que si on menait les deux études séparément : (A+) *versus* (A-) et (B+) *versus* (B-). L'autre avantage est que l'on peut comparer les différents traitements entre eux ainsi que l'influence qu'ont les différents traitements les uns sur les autres (interactions). En revanche, si l'objectif principal est de démontrer que le traitement A associé au traitement B est significativement plus efficace que chaque traitement pris individuellement, il faut inclure davantage de patients dans le groupe 4 (A+B+). L'analyse statistique d'un plan factoriel se fait en deux étapes. La première étape est la recherche d'une interaction. En l'absence d'interaction, l'effet des traitements est recherché en ajustant sur l'autre traitement [9].

Tableau II. Exemple d'un plan factoriel 2 x 2.

	Contrôle de A (A-)	**Traitement A (A+)**
Contrôle de B (B-)	Groupe 1	Groupe 2
Traitement B (B+)	Groupe 3	Groupe 4

Essai à unités de randomisation collectives

Principe

Un essai clinique randomisé par grappes *(cluster randomization trial)* est un essai dans lequel on ne randomise pas individuellement des sujets, mais des groupes de sujets qu'on appelle des « grappes » *(clusters)*. Ces unités de randomisation peuvent être des hôpitaux, des médecins, des familles, des villages, autant d'unités pour lesquelles les sujets qui composent une unité ne peuvent être considérés comme indépendants les uns des autres.

Explications

Ce type de randomisation permet d'éviter les contaminations potentielles entre patients dans les études cliniques comportementales ou pour l'évaluation des pratiques. Ce type d'étude est parfois le seul possible si le critère de jugement se mesure au niveau du groupe. L'analyse de ce type de schéma d'étude est particulière [10, 11].

Le nombre de sujets nécessaires

Pour pouvoir mettre en évidence l'efficacité ou la supériorité d'un traitement, il faut s'assurer que les résultats de l'essai, basés sur les analyses statistiques, ont peu de chances d'être dus au hasard.

Pour commencer, il faut définir le « bon » critère de jugement (survie globale, survie sans récidive, taux de PSA, survie sans progression, douleur, qualité de vie, etc.). Tous ces critères (résumés sous un format binaire, continu ou de type « survie ») peuvent avoir un intérêt, car ils vont être la base de la mesure de l'effet du groupe expérimental *versus* le groupe contrôle.

Il faut ensuite préciser la taille de la « différence » ou écart entre les groupes que l'on veut déceler avec le critère choisi – on parle de la taille de l'effet du traitement qui peut être une différence ou un rapport. Par exemple, si avec le traitement de référence A la probabilité de survie à 5 ans est de 60 % et que l'on veuille obtenir 75 % de survie avec le nouveau traitement B, la différence absolue à mettre en évidence sera de 15 %, alors que l'effet en termes de survie se mesure par un rapport de risque *(Hazard Ratio)*, que l'on peut estimer, dans le cas d'une survie exponentielle, par $\text{Log}(0,75) / \text{Log}(0,60) = 0,56$ et qui se traduit par une diminution du risque de 44 %. Pour détecter des différences de survie, on raisonne en nombre total d'événements (décès, récidives, métastases selon le critère de survie) à partir duquel on calcule un nombre total de patients à inclure dans un laps de temps raisonnable [7].

Il est également important de préciser dans quel sens on recherche cette différence ; dans l'exemple ci-dessus, on cherche à savoir si B est supérieur à A ou non (comparaison dite unilatérale) mais il y a d'autres cas où l'on cherchera à savoir si B est différent de A ou non, sans idée *a priori* sur le sens de la différence (comparaison dite bilatérale), le nombre de sujets nécessaires étant plus élevé dans le second cas.

Pour un critère continu, il faut avoir une idée de la dispersion des résultats dans la population étudiée, appelée « variance ». Elle doit s'estimer par des connaissances antérieures telles les résultats des essais de phase II.

Il faut également fixer les risques d'erreurs : le risque de première espèce β (la probabilité de trouver une différence à tort alors qu'il n'y en a pas), généralement égale à 5 %, et le risque de deuxième espèce *β* (la probabilité de ne pas déceler la différence alors que le traitement est efficace), généralement ≥ 80 % (*cf.* chapitre I.3 « Compréhension des tests statistiques », page 20).

C'est à partir de ces éléments que l'on peut calculer le nombre de sujets nécessaires, un nombre d'autant plus important que la différence que l'on veut déceler est petite, que la dispersion des résultats est grande et que les risques d'erreur choisis sont faibles.

Lors de la randomisation, on peut choisir de **stratifier** les patients pour être certain d'avoir des groupes comparables sur des facteurs pronostiques déjà connus ; les facteurs de stratification le plus souvent utilisés sont : le sexe, l'âge, la gravité de la maladie, le centre de soins... (*cf.* chapitre VI.2, page 344).

À retenir

La détermination du nombre de sujets dépend des éléments suivants :
– une estimation réaliste de l'effet du traitement attendu dans le groupe contrôle ;
– une estimation réaliste de l'effet du traitement espéré dans le groupe expérimental (différence ou rapport) ;
– le choix du risque alpha (en général bilatéral à 0,05 sauf pour les essais de non-infériorité pour lesquels certains prônent un risque unilatéral de 0,025) et du risque bêta (qui prend souvent les valeurs 0,10 ou 0,20) ;
– une estimation réaliste de taux d'inclusion attendu et de la durée d'inclusion (en général, la durée n'excède pas 5 ans) ;
– la durée de suivi de chaque patient dans l'étude.

Les essais de non-infériorité nécessitent d'inclure beaucoup plus de patients que les essais de supériorité puisque la marge d'équivalence est toujours plus petite que l'effet mis en évidence dans les essais de supériorité (qui doivent avoir précédé l'essai de non-infériorité).

Les aspects de méthodes et analyses

La population

Un essai randomisé est justifié quand l'état des connaissances, au moment où la participation dans l'essai est proposée aux individus, **ne permet pas de décider s'il est préférable, *a priori*, d'être exposé au nouveau traitement ou au traitement de référence (ou son placebo)**. Ce principe doit être acceptable pour le patient, le clinicien et l'investigateur qui propose l'essai.

Deux attitudes différentes peuvent être adoptées pour définir les critères d'inclusion des patients dans l'essai. La première attitude peut être de cibler certains patients et de n'inclure qu'une fraction de ceux-ci pour avoir une population aussi homogène que possible. Il peut aussi être pertinent d'identifier, dans ce contexte, un bénéfice dans un sous-groupe de patients avec des caractéristiques très précises. La seconde attitude, à l'opposé, est de prendre une population plus large et hétérogène qui permettra aussi un recrutement rapide et une généralisation plus aisée des résultats à l'ensemble des patients. Le fait d'être plus ou moins restrictifs sur les critères d'inclusion dépend ainsi de la question scientifique et des données acquises sur le traitement au moment de la planification de l'essai.

Le déroulement initial et le suivi d'un essai de phase III

La randomisation construit deux groupes initialement comparables. Encore faut-il que cette comparabilité soit maintenue au cours de l'essai et que rien ne vienne l'altérer durant le suivi *(follow-up)* ou lors de la mesure du critère de jugement, pour que ces deux groupes ne diffèrent tout au long de l'essai que par la nature du traitement qu'ils reçoivent. L'influence des différents facteurs de confusion doit donc être maintenue identique entre les deux groupes.

Par exemple si, au cours du suivi, les patients d'un groupe reçoivent plus de traitements concomitants que ceux de l'autre groupe, il sera impossible de savoir si la différence obtenue en fin d'essai est bien due au traitement étudié ou si elle provient de la différence entre les traitements concomitants. Une erreur systématique (biais) est aussi possible si la mesure du critère de jugement ne s'effectue pas de la même façon dans les groupes. On parle de biais de suivi (ou de réalisation) et de biais de mesure *(figure 1)*.

Déroulement initial

Randomisation

Le tirage au sort supprime les biais de sélection, équilibre la répartition des facteurs pronostiques (connus ou inconnus) entre les groupes comparés, garantit la validité des tests statistiques utilisés pour comparer les résultats observés, neutralise les « effets cohortes » *(cf. infra)*, égalise les données manquantes et permet l'utilisation de placebo.

• *Biais de sélection :* un nouveau traitement présentant d'importants effets secondaires sera plutôt réservé aux malades susceptibles de mieux le supporter ou à ceux pour lesquels l'espoir de guérison est faible. Il faut donc le comparer au même type de malades, ne recevant pas ce nouveau traitement, et non à ceux que l'on a « écartés » de cette prescription.

• *Répartition des facteurs pronostiques (connus ou inconnus) entre les groupes comparés :* par exemple, si l'on réalise un essai de chimiothérapie des cancers du sein dans lequel il n'est pas prévu de chirurgie ganglionnaire, on peut considérer que, grâce au tirage au sort, les malades porteurs de ganglions histologiquement envahis (facteur pronostique majeur non mesurable sans chirurgie) seront répartis équitablement entre les groupes comparés, certitude infiniment plus improbable si le groupe de référence est de type historique (ceux traités dans une période antérieure) ou bibliographique.

• *Validité des tests statistiques utilisés pour comparer les résultats observés :* la distribution et la variance des paramètres étudiés étant de ce fait « identiques » en moyenne dans les groupes comparés, les tests, permettant de dissocier des différences réelles des différences aléatoires, peuvent alors s'appliquer.

• *Neutralisation des effets cohortes :* les maladies évoluent au cours du temps pour de nombreuses raisons (changement dans les modes de prise en charge, dans les moyens diagnostiques, dans les traitements, etc.), et c'est tout cet ensemble qui concourt à l'amélioration des résultats que l'on peut constater en comparant deux groupes de malades traités à des périodes différentes, et non pas seulement la seule modification des traitements.

• *Répartition équilibrée des données manquantes :* le tirage au sort uniformise la collecte des données dans les groupes comparés. En effet, s'il existe un déséquilibre du nombre des perdus de vue, par exemple entre les groupes étudiés, selon que les perdus de vue sont des échecs ou des succès, les résultats de la comparaison peuvent varier. Or, dans les témoins historiques, ce nombre est toujours inconnu et donc source possible de conclusion erronée.

• *Placebo :* le tirage au sort permet l'utilisation d'un placebo dans les essais nécessitant la technique du double aveugle, technique souvent indispensable pour différencier l'effet pharmacologique de l'effet psychologique.

Aveugle

• **Définition :** il s'agit de la non-connaissance du traitement administré, *i.e.* non-connaissance du résultat de la randomisation (attribution du type de traitement à chaque individu).

• **Justification :** la conviction *a priori*, avouée ou non, que le traitement sera (ou ne sera pas) efficace et bien toléré risque d'influencer le comportement et le jugement de l'observateur et de l'observé : pour le suivi (observance au traitement de l'étude, consommation de traitements concomitants, présence aux visites de suivi) comme pour l'évaluation *(outcome assessor)*.

• **Le double insu** *(double blind)* pour lequel patient et évaluateur (souvent le médecin) ne connaissent pas le traitement administré évite toute différence dans le suivi et l'évaluation des deux groupes. Le principe du double insu consiste à faire que tous les patients, quelle que soit leur

appartenance à l'un des groupes de l'essai, apparaissent identiques. Cela est obtenu en ne révélant pas la nature exacte du traitement reçu par les patients. Les patients du groupe traité reçoivent un traitement strictement identique en apparence à celui reçu par les patients du groupe contrôle. Les patients du groupe traité sont donc indiscernables des patients du groupe contrôle.

Qui peut être en aveugle dans un essai randomisé ?

- Simple insu : le patient.
- Double insu : le clinicien en charge du suivi (*care provider*) ou l'évaluateur (*outcome assessor*).
- Triple aveugle : le biostatisticien qui fait les analyses des traitements évalués.

Placebo (du latin « je plairai »)

• **Définition :** produit qui a la même forme, la même couleur, la même odeur que celles du médicament étudié mais qui ne contient pas de substance active. La comparaison de son effet avec le médicament étudié permet de confirmer l'efficacité de ce dernier.

À distinguer de l'effet placebo (effet psychophysiologique bénéfique obtenu par l'administration du placebo : comprimés, liquides, injection ou procédures).

• **Principe :** l'effet placebo étant considéré comme le niveau plancher de l'efficacité (effet pharmacologique nul), l'efficacité d'un médicament est donc définie par sa supériorité par rapport au placebo – on devrait dire par rapport à « son » placebo. En effet, la rigueur scientifique impose que cet essai, pour être probant, se déroule dans des conditions telles que ni les patients ni les soignants ne savent quel patient prend lequel des deux produits comparés, le verum ou le placebo. Une telle procédure, appelée double aveugle, implique que le nouveau médicament et son placebo soient indiscernables l'un de l'autre. Par ailleurs, la prise en compte de l'effet nocebo (symptôme ou modification physiologique indésirable induite par la présence d'un placebo) est indispensable pour apprécier le profil de tolérance des produits actifs testés, l'ensemble des effets indésirables d'un médicament procédant à la fois de manifestations spécifiques en relation avec l'activité pharmacologique du produit et de manifestations non spécifiques en rapport avec le seul fait que le patient sait qu'il prend un médicament. Le placebo est donc non seulement utile à l'évaluation de l'efficacité des futurs médicaments mais aussi à celle de leur tolérance.

En cancérologie : des éléments pour comprendre un placebo

- Première description de l'effet placebo pour une régression spontanée d'un patient atteint de lymphome non hodgkinien. L'effet placebo existe donc en cancérologie mais on utilise plus fréquemment aujourd'hui, au vu de la gravité de la maladie, le traitement de référence plutôt qu'un placebo pour comparer l'efficacité.
- Utilisation d'un placebo surtout dans les symptômes fonctionnels (par ex. la douleur) ou devant l'absence de traitement de référence ou en combinaison (le traitement de référence et le traitement à évaluer *vs* le traitement de référence et un placebo).
- Il existe toujours un débat sur l'aspect éthique et scientifique du placebo quelle que soit la maladie.

Suivi de l'essai et maintien de la comparabilité

Enfin, le tirage au sort, s'il est dans la plupart des cas nécessaire, n'est pas suffisant pour assurer la validité d'un essai, car s'il sert à créer des groupes *a priori* comparables, encore faut-il que ces groupes le restent jusque et y compris pendant l'analyse des résultats, ce qui représente la deuxième difficulté de l'expérimentation thérapeutique. En effet, au cours du déroulement de l'essai, nombreuses sont les occasions de modifier unilatéralement les groupes de traitement : thérapeutiques complémentaires, surveillance à espace irrégulier, disparition des sujets… L'ensemble de ces phénomènes peut aboutir à totalement modifier la composition de l'un ou de l'autre groupe, et de nouveau la comparaison finale ne sera plus le seul fruit de la différence de traitement…

Le meilleur moyen pour éviter ce problème est de mener l'essai en « double aveugle » (ou double insu), c'est-à-dire que les traitements, objets de la comparaison, soient indiscernables pour le malade comme pour le médecin. Cette technique permet d'avoir un comportement tout à fait identique vis-à-vis des malades tout au long de l'essai quel que soit le groupe auquel ils appartiennent.

Pour des raisons éthiques ou matérielles, il est cependant parfois impossible de suivre une telle technique, il faut alors éviter au maximum toute déviation au protocole, suivre tous les malades entrés dans l'essai *sans exception*, avoir des critères de jugement aussi objectifs que possible ou faire évaluer les résultats par un observateur ignorant le traitement administré (lecture aveugle). Enfin, et toujours dans le but de ne juger que la différence liée aux effets du traitement, l'analyse des résultats doit *a priori* porter sur *tous* les sujets *inclus* au départ.

Si quelques exclusions s'avèrent inéluctables (absence de données concernant les sujets), elles doivent être clairement explicitées. Il est bien évident que plus il y aura de données manquantes ou d'écarts au protocole, plus la validité des résultats sera contestable.

Ainsi, la valeur scientifique d'un essai contrôlé dépend autant de la qualité du suivi des malades et des données recueillies que de la stratégie de départ. Une analyse statistique mal conduite pourra toujours être refaite, des données manquantes peuvent parfois être récupérées, mais un protocole mal suivi ou une stratégie inadéquate est une erreur fatale quant à la valeur des conclusions.

Or il ne faut pas oublier que nous menons ces essais thérapeutiques sur des êtres humains et que seule une étude scientifique correcte aboutissant à des conclusions extrapolables à d'autres patients, et se justifiant par ce fait, peut se défendre sur le plan éthique.

Finalement, il faut surveiller périodiquement le bon déroulement de l'essai : rythme des inclusions, description des exclus et inclus, exécution correcte du tirage au sort, respect du protocole en ce qui concerne les examens à faire, les traitements…, étude des abandons de traitement – les sujets qui ont reçu un traitement incomplet doivent être suivis comme les autres, en particulier en ce qui concerne le résultat final.

Les critères de jugement

Ces essais permettent de changer de traitement de référence et donc de changer les pratiques si le traitement évalué (ou la stratégie thérapeutique) est démontré comme supérieur en termes de bénéfice clinique. Les critères de jugement de ce bénéfice sont ainsi souvent, en cancérologie, la survie (*cf.* chapitre III.1 « Données de survie », page 129) ou la qualité de vie (*cf.* chapitre II.4 « Critères de qualité de vie relatifs à la santé », page 99).

Analyses

Les analyses des essais cliniques sont une étape cruciale qui doit suivre les recommandations internationales et pour lesquelles un comité de surveillance peut être une nécessité surtout lors des analyses intermédiaires (*cf.* chapitre VI.3, page 354).

Analyses intermédiaires

Une analyse intermédiaire est une analyse qui est faite sur le critère principal de l'essai avant la date prévue de l'analyse finale, surtout dans les essais qui prévoient l'inclusion d'un grand nombre de sujets sur une longue période. Elles sont de plus en plus souvent planifiées dans les protocoles, à partir des recommandations internationales (*International Conference on Harmonisation*, *cf.* chapitre VI.3, page 354), qui sont fondées sur des considérations à la fois éthiques et scientifiques, tout en garantissant la plus grande objectivité. Ces recommandations prévoient la mise en place d'un comité de surveillance indépendant qui peut recommander l'arrêt de l'essai dans les situations suivantes : intolérance trop élevée, efficacité hautement significative ou efficacité nulle. D'autres situations peuvent conduire à arrêter l'essai : rythme des inclusions trop lent, problèmes logistiques sévères, qualité des données médiocres ou considérations éthiques [12].

Après la fin des inclusions, une analyse intermédiaire permet de porter un jugement sur l'efficacité ou l'inefficacité du nouveau traitement par rapport au traitement de référence. Si tous les patients ont déjà été inclus mais que le nombre total d'événements attendus (celui qui a été calculé au début de l'essai pour assurer une puissance suffisante) n'est pas encore observé, une réponse positive précoce à la question posée permet de transmettre ces résultats à la communauté scientifique plus tôt que prévu, par une communication dans un congrès ou par la publication d'un article convaincant.

Pour construire un essai avec des analyses intermédiaires, tous les éléments habituellement demandés pour les essais avec une analyse finale unique doivent être renseignés. Les trois éléments supplémentaires sont : le type d'arrêt envisagé (supériorité, infériorité, futilité), le nombre d'analyses intermédiaires prévues, ainsi que la vitesse de dépense du risque α. Ce dernier point est assez technique mais a été développé dans le but de pouvoir proposer des plans d'expérience qui encadrent des analyses sauvages, car celles-ci peuvent donner des résultats très instables, surtout au début de l'essai [13].

Analyses finales

Dans tous les essais de phase III, l'analyse principale repose sur la population dite en « **intention de traiter** » (ITT ou FAS pour *Full Analysis Set*) pour pouvoir conclure à un effet du traitement testé « dans la vraie vie ». Toutefois, on peut également s'intéresser à l'analyse sur la population traitée selon le protocole, dite « **per-protocole** » (PP) qui tient compte des sorties d'étude, de l'**observance** du traitement et des éventuelles déviations au protocole. Ces analyses permettent d'évaluer la robustesse des analyses.

Parfois, des analyses de sous-groupes, de préférence prévues dans le protocole, permettent de préciser l'effet du traitement dans des populations sélectionnées par rapport à un facteur pronostique par exemple. La présentation de ces résultats sous forme de *Forest plots* (figure inspirée de la présentation des résultats d'une méta-analyse, *cf.* chapitre IV.5 « Méta-analyse d'essais randomisés », page 244) permet de visualiser l'effet du traitement dans chaque sous-groupe [14]. En revanche, ces résultats sont à prendre avec beaucoup de précaution, car tout comme les analyses intermédiaires, ils sont réalisés sur un échantillon plus petit, sont moins puissants et peuvent donc générer beaucoup de faux positifs [15]. Néanmoins, les analyses en sous-groupes peuvent servir à générer des hypothèses scientifiques pour faire une autre étude si la question est encore pertinente pour ce sous-groupe particulier.

À retenir

- **Analyse en ITT :** l'analyse principale compare les groupes tels qu'issus de la randomisation c'est-à-dire que tous les patients randomisés sont analysés dans le groupe dans lequel ils ont été randomisés, et l'analyse porte sur tous les patients inclus même si des déviations au protocole sont observées (inclusions à tort, mauvaise adhérence ou arrêt du traitement et perdus de vue). Enfin, tous les patients sont suivis jusqu'à la fin de l'étude qu'ils prennent ou non le traitement. L'analyse en ITT est la règle dans les essais cliniques visant à détecter une différence car elle diminue le risque de trouver cette différence par hasard.
- **Analyse en PP :** cette analyse ne prend en compte que les patients inclus qui étaient conformes aux critères d'inclusion (population dite éligible) et qui ont participé à l'essai du début à la fin, en étant strictement conformes au protocole de l'essai. Néanmoins, la définition de cette « conformité » peut être assez subjective et ne concerne donc qu'une population particulière de patients. Cette sorte de « sélection » des patients analysés favorise la détection d'une différence entre les traitements dans un essai contrôlé. C'est pourquoi cette analyse statistique n'est pas l'analyse principale de l'essai de supériorité tandis qu'elle est recommandée en première approche dans un essai de non-infériorité.

Les analyses de tolérance sont basées sur la population des patients randomisés qui ont commencé le traitement. Ils sont analysés dans le groupe du traitement réellement reçu, qui peut parfois être différent du groupe de randomisation.

Les essais de non-infériorité

Dans certains essais cliniques, il peut être intéressant de démontrer qu'un nouveau traitement fait « tout aussi bien » que le traitement de référence. C'est le cas des essais de désescalade thérapeutique, qui tentent de montrer une non-infériorité.

Comme il est impossible de démontrer une équivalence, ce qui reviendrait à accepter l'hypothèse nulle, un test statistique unilatéral est construit pour tester l'hypothèse nulle que le groupe contrôle est plus efficace que le traitement expérimental avec une marge acceptable [16]. Cette situation permet de rejeter l'hypothèse nulle, ce qui est la règle d'or dans la construction des tests statistiques. Des procédures fondées sur les intervalles de confiance (IC) peuvent aussi être utilisées [8]. Elles doivent exclure une non-infériorité avec une probabilité élevée [17]. La grande difficulté réside dans le choix de la marge de non-infériorité et dans sa justification clinique et statistique [18].

La *figure 2* présente sept situations correspondant aux résultats que l'on peut observer dans un essai de non-infériorité entre deux traitements, contrôle (C) et nouveau (T). Les effets en faveur de T sont à gauche de la ligne zéro et les effets en faveur du C sont à droite de la ligne zéro. La marge de non-infériorité a été fixée à 2 *(figure 2)*.

La non-infériorité est démontrée dans les situations 1, 4, 5 et 6 car la limite supérieure de l'IC à 95 % est inférieure à la marge de non-infériorité, quelle que soit l'estimation de l'effet *(tableau III)*.

Dans les situations où la marge est contenue dans l'IC, on ne peut conclure à la non-infériorité (cas 2, 3 et 7). Dans la situation 5, non seulement la non-infériorité est démontrée, mais la supériorité également, car la limite supérieure est inférieure à la valeur nulle. La situation 6 peut présenter des difficultés d'interprétation, car la non-infériorité est démontrée malgré un effet supérieur du groupe contrôle, mais cette situation est rare. Le cas 7 représente une situation ambiguë où la non-infériorité n'est pas démontrée (marge contenu dans IC) malgré une supériorité apparente de bras contrôle (borne inférieure supérieure à zéro) et une estimation de l'effet supérieure à la marge.

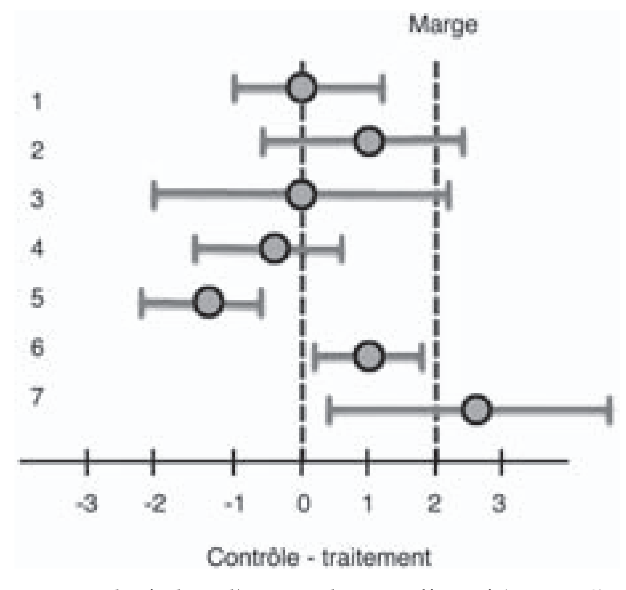


Figure 2. Exemple de six types de résultats d'un essai de non-infériorité (marge = 2).

Tableau III. Interprétation des sept situations de la figure 2.			
Situation	**Estimation C - T**	**Intervalle de confiance**	**Interprétation**
1	Effet semblable	Limite supérieure < 2	Non-infériorité démontrée
2	Effet favorise C	Limite supérieure > 2	Non-infériorité non démontrée
3	Effet semblable	Limite supérieure > 2	Non-infériorité non démontrée
4	Effet favorise T	Limite supérieure < 2	Non-infériorité démontrée Supériorité non démontrée
5	Effet favorise T	Limite supérieure < 2 Limite inférieure > 0	Non-infériorité démontrée Supériorité de T démontrée
6	Effet favorise C	Limite supérieure < 2 Limite inférieure > 0	Non-infériorité démontrée Supériorité de C démontrée
7	Effet favorise C	Limite supérieure > 2 Limite inférieure > 0	Non-infériorité non démontrée Supériorité de C démontrée

Essai de supériorité et de non-infériorité

Dans certains cas, les résultats d'un essai de supériorité peuvent suggérer une non-infériorité entre les traitements ou les résultats d'un essai de non-infériorité peuvent suggérer une supériorité du traitement évalué (cas 5). Un essai de non-infériorité peut permettre de conclure à une supériorité si la borne supérieure de l'IC à 95 % de la différence se trouve entièrement au-dessus de zéro.

Mais un essai de supériorité ne peut généralement pas mettre en évidence une non-infériorité sauf si une marge de non-infériorité a été définie dans le protocole et que l'essai a été réalisé selon les standards stricts des essais de non-infériorité, ce que l'on appelle des essais hybrides [19].

La conclusion d'un essai

Si la différence est significative, il convient toujours de tenir compte du degré de signification et de l'IC de la différence et de rapprocher ces résultats des connaissances antérieures à l'essai. La conclusion a d'autant plus de pouvoir de conviction que la différence est très significative et concorde avec des hypothèses explicatives.

Enfin, les conclusions d'un essai doivent être considérées dans le contexte de l'ensemble des données disponibles sur le traitement à l'étude. Quand un nouveau traitement est bénéfique mais a un effet modéré, il faut s'attendre à voir publier des résultats apparemment contradictoires, certains essais étant significatifs et d'autres non. Seule une analyse globale de l'ensemble des données évaluant, sans biais, l'intérêt du traitement permet à la fois d'étudier l'hétérogénéité

Énoncé des conclusions

Pour pouvoir définir la portée de l'essai, il faut le conclure en donnant clairement les résultats et en reprenant les différents points du protocole : sujets, traitements, type d'essai, résultats. On écrira par ex. : N sujets ayant subi l'intervention X ont reçu après tirage au sort et dans des conditions doublement à l'aveugle : soit le produit A administré à telle dose, sous telle forme et pendant telle durée, soit un placebo administré dans les mêmes conditions.
Le produit A a montré une efficacité significativement supérieure à celle du placebo (par ex. : 40 % *vs* 20 %, $p = 0{,}05$) *ou bien* il n'a pas été mis en évidence de différence significative entre l'efficacité du produit A et celle du placebo (par ex. : 35 % *vs* 30 %, $p = 0{,}30$).
Lorsque l'on ne trouve pas de différence significative, il ne faut surtout pas conclure à l'équivalence des traitements. L'absence de différence perd encore plus de la valeur si le nombre de sujets n'est pas celui prévu au départ...

entre essais et surtout de tester l'intérêt du traitement de façon globale. Cette méthode, appelée méta-analyse (*cf.* chapitre IV.5, page 244), a été utilisée récemment pour évaluer l'efficacité de la chimiothérapie et de l'hormonothérapie adjuvantes dans les cancers du sein opérables d'emblée.

Pour finir

Avant la mise en place d'un essai de phase III, il faut s'assurer des conditions suivantes :
– l'essai ne peut être entrepris que s'il existe au départ un équilibre entre les avantages et les inconvénients éventuels des traitements que l'on compare – on parle d'ambivalence ou *equipoise*) ;
– l'essai doit être conduit selon une méthodologie scientifiquement rigoureuse dans un milieu médical compétent et bien équipé – les équipes comportent des professionnels spécialistes de la recherche clinique ;
– le consentement libre et éclairé du malade doit être obtenu – le dossier aura alors été validé par le comité de protection des personnes en France et par l'autorité réglementaire telle que l'Agence française de sécurité sanitaire des produits de santé (Afssaps).

La conférence internationale d'harmonisation des essais cliniques (ICH pour *International Conference on Harmonization*) a produit des recommandations adoptées en Amérique du Nord, en Europe et au Japon qui incluent des guides de bonnes pratiques cliniques (ICH Topic E6), des considérations générales sur les essais cliniques (ICH Topic E8), des recommandations pour les principes statistiques dans les essais cliniques (ICH Topic E9) et aussi un document guide pour le choix du groupe contrôle dans les essais (ICH Topic E10).

Par ailleurs, depuis 2001, sont édités par le groupe Consort différents textes pour rappeler les éléments nécessaires de précision dans les publications des essais cliniques (http ://www.consort-statement.org/ref).

Références

1. Fisher RA, MacKenzie WA. Studies in crop variation: II. The manurial response of different potato varieties. *J Agric Sci* 1923 ; 13 : 311.

2. Amberson JB JR, McMahon BT, Pinner M. A clinical trial of sanocrysin in pulmonary tuberculosis. *Am Rev Tuberc* 1931 ; 24 : 401.

3. Diehl HS, Baker AB, Cowan DW. Cold vaccines: An evaluation based on a controlled study. *JAMA* 1938 ; 111 : 1168.

4. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis: A medical research council investigation. *Br Med J* 1948 ; 2 : 769.

5. Cucherat M, Lievre M, Leizorovicz A, Boissel JP. *Lecture critique et interprétation des résultats d'essais cliniques pour la pratique médicale.* Paris : Médecine Science Flammarion, 2004, 256 pages.

6. Buyse ME, Staquet MJ, Sylvester RJ. Cancer clinical trials: Methods and practice. Oxford: Oxford Medical publication, 1984, 481 pages.

7. Sylvester R, Van Glabbeke M, Collette L, *et al*. Statistical methodology of phase III cancer clinical trials: Advances and future perspectives. *Eur J Cancer* 2002 ; 38 : S162-S168.

8. Elie C, de Rycke Y, Jais JP, *et al*. Aspects méthodologiques et statistiques des essais d'équivalence et de non-infériorité. *Rev Epid Sante Pub* 2008 ; 56 : 267-77.

9. Green S. Factorial design considerations. *J Clin Oncol* 2002 ; 20 (16) : 3424-30.

10. Campbell MK, Elbourne DH, Altman DG CONSORT group. CONSORT statement extension to cluster randomised trials. *BMJ* 2004 : 326 (7441) : 702-8.

11. Giraudeau B, Ravaud P. Preventing bias in cluster randomosed trials. *PLoS Med* 2009 : 6 (5) : e1000065.

12. Kramar A, Paoletti X. Analyses intermédiaires. *Bull Cancer* 2007 ; 94 (11) : 965-74.

13. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979 ; 35 : 549-56.

14. Pocock SJ, Travison TG, Wruck LM. How to interpret figures in reports of clinical trials. *BMJ* 2008 ; 336 (7654) : 1166-9.

15. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine – Reporting of subgroup analyses in clinical trials. *N Engl J Med* 2007 ; 357 : 2189-94.

16. Blackwelder WC. "Proving the null hypothesis" in clinical trials. *Control Clin Trials* 1982 ; 3 : 345-53.

17. Laster LL, Mary F, Johnson MF. Non-inferiority trials: The 'at least as good as' criterion. *Statist Med* 2003 ; 22 : 187-200.

18. Ng TH. Noninferiority hypotheses and choice of noninferiority margin. *Statist Med* 2008 ; 27 : 5392-406.

19. Freidlin B, Korn EL, George SL, Gray R. Randomized Clinical trials for assessing noninferiority when superiority is expected. *J Clin Oncol* 2007 : 5019-23.

Essais cliniques de phase 0 en cancérologie

M.Q. Picat, N. Houédé, E. Chamorey

Le développement de nouveaux médicaments est un processus complexe, long et coûteux. La séquence conventionnelle d'études précliniques chez l'animal, puis cliniques chez l'homme peut durer de 10 à 15 ans avec des coûts qui ne cessent d'augmenter [1]. Actuellement, un nouveau médicament entrant dans un processus de phase I n'a que 8 % de chances d'intégrer le marché du médicament alors que ce taux était de 15 % en 1985.

Pour les molécules de lutte contre le cancer, ce taux de succès n'est que de 5 % [2]. Les échecs ont souvent lieu en phase III, en rapport avec un ratio bénéfice/risque défavorable ou à un apport bénéfique trop faible par rapport aux traitements existants. Ces échecs tardifs, survenant dans les dernières phases de développement du médicament, entraînent une perte de temps et de ressources non négligeables. De plus, l'application des méthodes classiques aux molécules dites ciblées pose de réelles difficultés. Les échecs en phase I sont précoces liés à une impossibilité de déterminer la dose recommandée. En effet, le suivi des seules toxicités lors de l'escalade de dose en phase I, à la recherche de la dose maximale tolérée, ne permet plus de définir la dose et le schéma d'administration optimaux de ces nouvelles molécules.

Si les avancées scientifiques permettent d'accroître le nombre de nouveaux médicaments potentiels, la pression économique de plus en plus exacerbée nécessite le développement d'outils adaptés pour rejeter le plus rapidement possible ceux qui seraient inefficaces ou toxiques.

Contexte de mise en place

Dans ce contexte, la *Food and Drug Administration* des États-Unis (FDA) et l'Agence européenne du médicament (EMEA) ont souligné le besoin urgent de développer de nouveaux outils destinés à faire le lien entre les découvertes de nouvelles molécules par les scientifiques et les premières applications médicales [3, 4]. Cette volonté de promouvoir et d'accélérer le développement de nouvelles molécules a conduit à l'élaboration de deux guides :

- *Guidance for Industry, Investigators, and Reviewers : Exploratory Investigational New Drug* [2] ;
- *Guideline on strategies to identify and mitigate risks for first-in-human clinical trials with investigational medicinal products* [5].

Ces recommandations avaient pour objectif de répondre précocement à des questions précises permettant de décider ou non la poursuite du développement de la molécule. Elles préconisent ainsi le développement d'études cliniques exploratoires dites de phase 0. Il s'agit d'un nouveau type d'essai clinique, très précoce, conduit sur un nombre réduit de patients (en principe moins de 30), avec des durées de traitement courtes (en principe moins de 7 jours). Pour éviter la survenue de toxicité, les doses administrées de la molécule à l'étude sont faibles [6]. Elles sont toujours inférieures à la dose sans effet toxique *(No Observable Adverse Effect Level)*, dose la plus élevée d'une substance pour laquelle aucun effet toxique n'est observé, souvent utilisée pour le calcul de base de la première dose d'administration chez l'homme.

Ces études très spécifiques de phase 0, actuellement menées aux États-Unis, se situent donc entre les études de toxicologie animale et le classique essai de phase I de recherche de toxicologie humaine. Il s'agit théoriquement d'une première administration d'une nouvelle molécule chez l'homme, les patients étant soumis à un ensemble de tests (paramètres pharmacologiques, biomarqueurs, biopsies, imageries, etc.). Le but est d'établir la preuve du concept d'efficacité biologique de la molécule avant ses premiers tests en phase I de toxicologie et de recherche de doses.

Les essais de phase 0 peuvent également être utiles pour préciser les mécanismes d'action, de toxicité, d'interaction ou de synergie entre composés ayant déjà été testés sur l'homme.

D'un point de vue de l'industrie pharmaceutique, ces essais de phase 0 présenteraient l'avantage de pouvoir sélectionner plus rapidement les molécules, avec peu de patients et une quantité limitée de principe actif [7].

Fin 2009, presque une trentaine d'articles relatifs à ces essais a été publiée (1 article en 2006, 5 en 2007, 11 en 2008 et 9 en 2009). Les champs concernés par la publication d'articles sur les essais de phase 0 sont la cancérologie et la pharmacologie. Les principales revues de publication sont en particulier le *Clinical Cancer Research*, le *Journal of Clinical Oncology* et l'*European Journal of Cancer*.

Le concept d'essai clinique de phase 0 est donc un concept récent. Dans la littérature, ces essais sont nommés sous différentes appellations : *« phase 0 (clinical) trial(s) »*, *« phase 0 study/studies »*, *« phase 0 first-in-human clinical trial »*, *« exploratory trial »*, *« early clinical trial »*. En 2009, il n'existe pas encore de *Mesh term* référencé pour les essais de phase 0 dans le moteur de recherche PubMed.

Apports et objectifs

Objectifs

L'intérêt potentiel des essais de phase 0 réside en particulier dans l'acquisition très précoce de données pharmacocinétiques et pharmacodynamiques. Elles permettent la justification du rationnel de l'étude en apportant des éléments dans le choix des doses initiales, sur les éventuels antagonismes ou sur les synergies d'action [8]. Les objectifs principaux des essais cliniques de phases 0 se déclinent en trois grandes catégories [9] :

- démontrer que les hypothèses précliniques sont validées (c'est-à-dire que le mécanisme d'action supposé chez l'animal est bien celui retrouvé chez l'homme) et ainsi décrire l'effet de la molécule à l'étude sur les cibles intratumorales ;

• sélectionner la molécule la plus intéressante chez l'homme en fonction de critères pharmacocinétiques et pharmacodynamiques objectifs. En effet, en règle générale, une molécule est sélectionnée par rapport à d'autres critères dès la phase préclinique qui est basée uniquement sur des résultats sur l'animal ou *in vitro*. Le caractère limité des modèles précliniques et parfois leur faible aptitude à reproduire un modèle humain entraînent des difficultés pour sélectionner le meilleur candidat ;
• évaluer un schéma d'administration optimal de la molécule ou une association thérapeutique en termes de synergie, d'antagonisme ou d'interaction potentielle, dans le but de justifier un rationnel pour une étude de phase I.

Enfin, il peut également être envisagé de développer les essais de phases 0 sur des molécules potentiellement efficaces, mais dont le développement aurait été stoppé du fait d'une mauvaise évaluation de la toxicité ou de l'efficacité initiale. Cela pourrait aider à mieux comprendre et maîtriser certains événements survenant plus tard dans le développement de la molécule (interactions médicamenteuses, pharmacocinétique chez des sujets à risque, mécanisme de toxicité, ciblage de population à traiter, etc.).

Essais de phase 0 *versus* essais de phase I

Approche des essais de phase I

Le but d'un essai clinique de phase I en oncologie est de définir avec précision la dose recommandée pour les futurs essais de phase II. Pour définir la dose recommandée, il convient d'estimer les toxicités dose limitantes, qui sont les différents types de toxicités limitant l'augmentation de dose du traitement, ainsi que la dose maximale tolérée, qui est la dose entraînant une toxicité dose limitante intolérable pour le patient, non gérable ou irréversible. En pratique, la dose recommandée pour les essais de phase II correspondra à la posologie qui précède la dose maximale tolérée [10].

Afin de définir la dose recommandée, la planification d'une étude de phase I de cancérologie s'articule autour de quatre principales étapes [11] :
• sélection de la dose initiale ayant une toxicité présumée acceptable ;
• choix des posologies pour les paliers de dose successifs ;
• choix d'un schéma d'augmentation de dose ;
• déroulement de l'étude, avec inclusion de cohortes de patients dans les paliers de dose successifs jusqu'à l'obtention de la dose maximale tolérée.

Approche exploratoire des essais de phase 0

Si les essais de phase I ont clairement un objectif axé sur la recherche de toxicité de la molécule, les essais de phase 0, quant à eux, n'ont pas d'objectif de recherche de toxicité, c'est pourquoi il est nécessaire de clairement distinguer ces deux types d'essais [12].

Dans les essais de phase 0, la première administration de la molécule chez l'homme à des doses faibles exclut les risques toxicologiques *a priori* et permet d'acquérir des données pharmacocinétiques et pharmacodynamiques précoces. Ces données sont nécessaires pour aider à la décision, notamment quand il s'agit de :

- stopper le développement de molécules non adaptées ;
- établir des critères de jugement biologiques prédictifs d'efficacité clinique ;
- rechercher précocement l'effet pour lequel la molécule est développée ;
- obtenir des données pertinentes sur la distribution de la molécule et de son effet sur la cible thérapeutique identifiée.

Les principales différences entre essais cliniques de phase 0 et de phase I sont présentées dans le *tableau I*.

Tableau I. Principales différences théoriques entre les essais cliniques de phase 0 et de phase I (d'après [19]).

	Essai clinique de phase I	**Essai clinique de phase 0**
Objectif principal	Établir une dose maximale tolérée	Établir l'effet de la molécule sur la cible d'intérêt
Sélection de dose	Sécurité d'emploi de la molécule et sa toxicité	Dose maximum utilisée toujours inférieure à la dose sans effet toxique *(No Observable Adverse Effect Level)*
Nombre de patients	Généralement > 30	Moins de 30 en principe
Doses administrées	Multiples	Limitées *(maximum 14 jours)*
Bénéfice thérapeutique	Non attendu ; cependant, la réponse tumorale peut être évaluée dans le cas où un bénéfice clinique serait retrouvé	Aucun
Biopsies tumorales	Optionnel	Réalisées pour évaluer l'effet de la molécule sur la cible
Pharmacocinétique-pharmacodynamie	Échantillons souvent analysés ultérieurement	Échantillons analysés en temps réel

Modalités de mise en œuvre

Cadre réglementaire

Les essais cliniques de phase 0 sont régis par les mêmes règles que tout essai clinique. Ils doivent se faire dans le respect des bonnes pratiques cliniques et être fondés sur des bases scientifiques solides et légitimées en accord avec les principes éthiques énoncés dans la Déclaration d'Helsinki.

Les procédures opérationnelles standards, qui définissent le rôle et les obligations des différents intervenants dans un essai clinique, doivent être respectées et clairement envisagées et rédigées à l'avance, de la collecte des échantillons jusqu'à la manière d'analyser les données [13].

Protocole d'étude, notes d'information et formulaire de consentement

Le projet d'essai clinique de phase 0 est mis en place selon un protocole défini mentionnant notamment le schéma d'étude, la population d'étude, le nombre de patients impliqués et la molécule concernée.

Une note d'information à l'attention du patient ainsi qu'un formulaire de consentement doivent être rédigés. L'équipe de recherche doit clairement renseigner les détails propres à l'étude de phase 0, tels que son objectif, sa durée, ses contraintes, les procédures requises et les examens potentiels à réaliser (biopsies par ex.). Le formulaire de consentement éclairé doit clairement préciser que les doses de la molécule administrée sont trop faibles pour entraîner *a priori* une toxicité, mais aussi un quelconque bénéfice thérapeutique [14].

Aspects pratiques : conditions à satisfaire pour conduire une phase 0

La mise en place d'une étude de phase 0 requiert une collaboration multidisciplinaire forte impliquant cliniciens, biologistes et pharmaciens. Ces collaborations devront permettre d'orienter la recherche et la mesure de paramètres objectifs et pertinents.

Pour obtenir des données d'une très grande qualité, le site de réalisation de l'essai doit être pourvu d'une capacité technique adaptée au moins identique à ceux nécessaires à la réalisation d'un essai de phase I (avec autorisation de lieu). Le type de tissu ou de liquide biologique prélevé doit être parfaitement défini, les conditions et sites de prélèvement standardisés ainsi que les conditions de conservation et de transport. Les biopsies de tissus ne pourront être envisagées que si les concentrations sanguines ou les dosages effectués sur les tissus de substitution ne permettent pas d'évaluer l'effet pharmacodynamique souhaité sur la cible tumorale. Les techniques analytiques d'identification et de dosages pharmacologiques doivent être parfaitement validées. Elles doivent être applicables à tout patient et apporter le plus d'informations possibles sur l'effet à mesurer tout en minimisant les contraintes pour le patient, le personnel de soins et les laboratoires [10].

Les analyses des données doivent être réalisées en temps réel et les résultats communiqués rapidement aux équipes cliniques afin de prendre une décision éclairée concernant l'intérêt de réaliser ou pas des prélèvements biopsiques invasifs. D'autre part, il est souhaitable d'avoir une idée sur la variabilité de l'effet pharmacodynamique en fonction de l'hétérogénéité tumorale, de la chronopharmacologie ou d'autres facteurs pouvant influencer l'évaluation de l'effet mesuré [15].

Un exemple d'étude de phase 0

On peut illustrer ce concept d'essai clinique précoce de phase 0 à travers l'expérience de l'équipe de Kummar. Des premiers travaux avaient été initiés en 2007 sur l'ABT-888, un inhibiteur de la poly(ADP-ribose) polymérase (PARP) et avaient été présentés à la conférence annuelle de l'*American Society of Clinical Oncology* [16]. Cette molécule avait été élaborée dans le but de contrer la résistance aux chimiothérapies. En inhibant le système de réparation des cassures simple brin d'ADN dans les cellules cancéreuses, elle permettrait de favoriser l'action d'autres agents anticancéreux. En 2009, Kummar *et al.* ont publié dans le *Journal of Clinical Oncology* une étude de phase 0, mise en œuvre sous la supervision de la FDA et destinée à définir le schéma d'administration de l'ABT-888 [17]. Les patients inclus étaient atteints de cancers ou d'hémopathies résistantes à au moins une ligne de chimiothérapie et avaient signé un consentement éclairé qui expliquait la finalité de l'étude ainsi que la nécessité de réaliser des biopsies. L'objectif de l'étude était de valider la pharmacocinétique de l'ABT-888 et de déterminer une dose et un schéma d'administration de la molécule permettant d'inhiber la PARP dans la tumeur et le sang. Pour cela, 5 niveaux de dose ont été testés (10, 25, 50, 100 et 150 mg en dose unique) avec 3 patients par niveau. La dose de départ, très faible, correspondait à 1/50e de la dose n'entraînant pas d'effets indésirables en phase préclinique chez le chien, espèce la plus sensible. Plusieurs prélèvements à des fins pharmacocinétiques et pharmacodynamiques, incluant 2 biopsies, étaient prévus dans le protocole. Le critère de jugement principal était une inhibition de la PARP d'au moins 50 % dans les 3 à 6 heures après administration de l'ABT-888. Ce n'est pas une nouvelle méthode statistique, mais après un gros travail de laboratoire, le biologiste a établi qu'une variation de 50 % de l'inhibition de PARP était biologiquement significative.

Vingt-quatre patients ont été inclus dans cette étude. Aucune toxicité significative n'a été identifiée. Une réduction significative de la PARP a été observée à la fois dans la tumeur et dans le sang aux doses de 25 et 50 mg pour 5 patients sur 6. Une biopsie réalisée 24 heures après administration chez 3 patients a révélé une réactivation de la PARP chez 2 patients. La recherche d'une corrélation entre la mesure de l'inhibition de la PARP dans la tumeur et dans le sang n'était pas significative même si une tendance se dégageait nettement. Enfin, les études de pharmacocinétique et de pharmacodynamie n'ont pas mis en évidence de corrélation statistiquement significative entre la concentration maximale du nouveau produit et l'inhibition de la PARP dans le sang, indépendamment de l'effet de la dose.

Les auteurs ont conclu avec enthousiasme à un succès. Ils ont pu définir, en moins de 6 mois, par une étude précoce des paramètres pharmacocinétiques et pharmacodynamiques, une dose d'ABT-888 biologiquement efficace sur la cible, à tester lors de la phase I ultérieure. La dose de départ recommandée pour les études de phase I de l'ABT-888 a été fixée à 10 mg 2 fois par jour, sur la base d'un rationnel scientifique puisque la PARP était inhibée à 90 % avec 25 mg et que sa réactivation était identifiable 24 heures après l'administration.

Ainsi, ce concept d'étude de phase 0 paraît pertinent, même si dans cet essai, certains résultats sont discutables (pas de relations pharmacocinétiques et pharmacodynamiques statistiquement significatives, ni de corrélation entre l'évolution de la PARP dans le sang et la tumeur). Néanmoins, il est admis que les résultats de cette étude seront probablement très importants pour la

suite du développement de cette molécule ABT-888, notamment ceux portant sur son rythme d'administration, puisqu'il a été posé un rationnel, *in vivo* chez l'homme, d'une administration biquotidienne. Sans cette étude, il est probable que le médicament aurait été testé une fois par jour dans le cadre d'une étude de phase I.

Point sur la situation en France

En novembre 2008 ont eu lieu les Rencontres nationales de pharmacologie clinique de Giens XXIV. À cette occasion, une table ronde s'est attachée à définir les essais cliniques de phase 0 et à proposer des recommandations relatives à leur mise en place en France (prérequis minimums acceptables en termes de données animales, de qualité pharmaceutique de la molécule expérimentale et des conditions de réalisation de l'essai). Les principales recommandations du groupe peuvent se résumer ainsi :

- un essai exploratoire de phase 0 est réalisé si c'est une étape justifiée, qui se fait dans le respect des prérequis et est utile dans le développement d'une molécule ;
- les populations dites vulnérables (sujets âgés et enfants) ne devraient pas être concernées par ce type d'essais, sauf exception dûment justifiée ;
- le dossier pharmaceutique tient compte de la courte durée de l'administration de la molécule et des faibles doses utilisées. Les règles de bonnes pratiques de fabrication peuvent être adaptées. Néanmoins, les qualités pharmaceutiques de la molécule (caractérisation de la substance et formulation minimale) doivent être prouvées ;
- le protocole doit comporter la réponse à un ensemble de questions pertinentes concernant la phase d'administration de la molécule, en particulier la justification sur les choix de la dose de départ et des éventuels paliers ;
- une présoumission à l'Agence française de sécurité sanitaire des produits de santé (Afssaps) est conseillée ainsi qu'une soumission au Comité de protection des personnes (CPP) pour garantir les droits des personnes et leur sécurité ainsi que la confidentialité des informations qui les concernent.

Au terme du débat, il est apparu que les essais cliniques de phase 0 sont possibles en France. Les experts ont particulièrement insisté sur la nécessité de s'assurer de la protection des personnes et de prendre des précautions pour éviter des débordements liés à la pression de l'industrie pharmaceutique.

Enjeux liés aux essais de phase 0

Éthique

Les essais de phase 0 posent un problème éthique puisque les patients ne peuvent espérer aucun bénéfice thérapeutique en contrepartie de leur participation à un protocole invasif incluant potentiellement des gestes invasifs de type biopsies [18, 19]. D'autre part, même si les doses administrées sont faibles, il existe toujours un risque pour le patient.

Un autre aspect éthique est également soulevé. Bien que la durée de participation à l'étude soit courte, la participation à un essai de phase 0 peut exclure le patient d'un autre protocole de recherche qui pourrait lui apporter un bénéfice thérapeutique potentiel.

Ainsi, dans le cadre de la mise en place d'un essai de phase 0, il est conseillé de prendre très précocement l'avis d'un CPP pour garantir au mieux la sécurité des personnes [20, 21].

Méthodologie statistique

L'analyse des essais de phase 0, impliquant un très petit nombre de patients et des effets limités, nécessite l'utilisation de méthodes statistiques adaptées à de petits effectifs [22]. La réponse pharmacodynamique considérée comme statistiquement significative pour un individu doit être définie au préalable. Les paramètres à prendre en compte sont en particulier le niveau basal du biomarqueur dans le tissu de substitution ou dans la tumeur (utilisé comme critère de jugement) avant l'administration de la molécule ainsi que la variation attendue du biomarqueur après administration de la molécule. La variation théorique attendue du biomarqueur est souvent définie à partir de la phase préclinique chez l'animal ou au terme d'un travail préalable considérable d'études de laboratoires *in vivo*.

Dans les travaux *princeps* sur l'ABT-888 présentés à l'*American Society of Clinical Oncology* en 2007 avec 3 patients par niveau de dose, les auteurs avaient considéré une réponse pharmacodynamique comme statistiquement significative au seuil alpha de 0,05 quand la PARP variait de plus de 2,3 fois l'écart-type de son niveau basal après administration de l'ABT-888 [9].

La gestion des difficultés liées au faible effectif de patients, aux effets limités des molécules à l'étude et à la variabilité intra-patient avant et après administration de la molécule à l'étude nécessite l'implication indispensable d'un statisticien pour l'analyse d'un essai de phase 0. Le statisticien doit être consulté dès la conception du schéma et de la mise en place de l'étude.

Questions en suspens

Au-delà de l'aspect éthique, des questions relatives à la mise en place de ces essais de phase 0 restent ouvertes. Quelles molécules sont réellement concernées par ces essais de phase 0 [9] ? En effet, les molécules à index thérapeutique trop étroit (cas des molécules cytotoxiques en cancérologie) ne peuvent être concernées par ces essais de phase 0 car on ne peut observer l'effet pharmacodynamique de la molécule en raison de l'administration de doses réduites.

D'autre part, est-on toujours certain de la validité des biomarqueurs utilisés en phase 0 pour la validation des molécules ? À défaut, quel est le risque d'écarter une molécule efficace à tort ? Ainsi, si la phase 0 est utile pour écarter les molécules inintéressantes, ne pourrait-elle pas retarder l'arrivée de molécules efficaces en clinique ? Enfin, on peut se demander si les essais de phase 0 vont réellement permettre d'aller plus vite dans le développement de nouvelles molécules que le processus traditionnel [23] ?

Conclusion

Les essais cliniques exploratoires de phase 0 constituent une perspective d'avenir pour le développement de nouveaux médicaments dit ciblés contre le cancer. La phase 0 peut être considérée comme le « chaînon manquant » qui permettrait de faire le lien entre la phase préclinique chez l'animal et la phase clinique chez l'homme. En s'immisçant entre ces deux phases, elle répond à de nouvelles préoccupations. Alors que la phase I vise à déterminer la toxicologie et la sécurité d'utilisation de la molécule par la recherche de la dose maximale tolérée notamment, la phase 0 vise à la compréhension de la pharmacodynamie et de la pharmacocinétique, sans ambition thérapeutique et, théoriquement, sans risque toxicologique puisque les doses administrées sont faibles.

Ainsi, l'intérêt de ces essais cliniques de phase 0 est double : tout d'abord scientifique, par l'acquisition de données précoces utiles à la conduite des phases I, puis économique, par la présélection des molécules à fort potentiel innovant. Néanmoins, ces essais sans bénéfice individuel posent des questions d'ordre éthique. Il est donc nécessaire d'établir des algorithmes de décision clairs pour optimiser l'application de ces essais de phase 0 [24], en gardant comme priorité la protection des personnes. Le meilleur moyen est de continuer à en discuter au cas par cas selon les molécules concernées. Les prochaines années permettront de positionner si ces essais de phase 0 seront d'utilité dans le développement de nouvelles molécules en cancérologie.

Références

1. Marchetti S, Schellens JHM. The impact of FDA and EMEA guidelines on drug development in relation to phase 0 trials. *Br J Cancer* 2007 ; 97 : 577-81.

2. US Food and Drug Administration. *Guidance for industry, investigators, and reviewers: Exploratory investigational new drug studies*. FDA, 2006 : http://www.fda.gov/cder/guidance/7086fnl.pdf.

3. Kinders R, Parchment RE, Ji J, *et al*. Phase 0 clinical trials in cancer drug developement: From FDA guidance to clinical practice. *Mol Interv* 2007 ; 7 : 325-34.

4. Schellens JHM. Phase 0 (zero) clinical trials: More than zero benefit? *EJC* 2009 ; 45 : 728-9.

5. European Medicines Agency. *Guideline on strategies to identify and mitigate risks for first-in-human clinical trials with investigational medicinal products*. EMEA, 2007: http://www.emea.europa.eu/pdfs/human/swp/2836707enfin.pdf.

6. Takimoto CH. Phase 0 clinical trials in oncology: A paradigm shift for early drug development? *Cancer chemoter Pharmacol* 2009 ; 63 : 703-9.

7. Eliopoulos H, Giranda V, Carr R, *et al*. Phase 0 trials: An industry perspective. *Clin Cancer Res* 2008 ; 14 : 3683-8.

8. Jacobson-Kram D, Mills G. Leveraging exploratory investigational new drug studies to accelerate drug development. *Clin Cancer Res* 2008 ; 14 : 3670-4.

9. Murgo AJ, Kummar S, Rubinstein L, *et al*. Designing phase 0 cancer clinical trials. *Clin Cancer Res* 2008 ; 14 (12) : 3675.

10. Chamorey E. *Méthodologie des essais de phase précoce en cancérologie : évolution des schémas et apport de la pharmacologie*. Thèse de Doctorat d'Université ; Université de la Méditerranée, 2009.

11. Calvert AH, Plummer R. The development of phase I cancer trial methodologies: The use of pharmacokinetic and pharmacodynamic end points sets the scene for phase 0 cancer clinical trials. *Clin Cancer Res* 2008 ; 14 : 3664-9.

12. Kummar S, Rubinstein L, Kinders R, *et al.* Phase 0 clinical trials: Conceptions and misconceptions. *Cancer J* 2008 ; 14 : 133-7.

13. Kummar S, Kinders R, Rubinstein L, *et al.* Compressing drug development timelines in oncology using phase "0" trials. *Nat Rev Cancer* 2007 ; 7 : 131-9.

14. Gutierrez M, Collyar D. Patient perspectives on phase 0 clinical trials. *Clin Cancer Res* 2008 ; 14 : 3689-91.

15. Doroshow JH, Parchment RE. Oncologic phase 0 trials incorporating clinical pharmacodynamics: From concept to patient. *Clin Cancer Res* 2008 ; 14 (12) : 3658-63.

16. Kummar S, Kinders R, Gutierrez M, *et al.* NCI phase 0 working group. Inhibition of poly (ADP-ribose) polymerase (PARP) by ABT-888 in patients with advanced malignancies: Results of a phase 0 trial. *J Clin Oncol* 2007 ; 25 : 3518.

17. Kummar S, Kinders R, Gutierrez M, *et al.* Phase 0 clinical trial of the poly (ADP-ribose) polymerase inhibitor ABT-888 in patients with advanced malignancies. *J Clin Oncol* 2009 ; 27 : 2705-11.

18. Hill TP. Phase 0 trials: Are they ethically challenged? *Clin Cancer Res* 2007 ; 13 (3) : 783-4.

19. Abdoler E, Taylor H, Wendler D. The ethics of phase 0 oncologytrials. *Clin Cancer Res* 2008 ; 14 : 3692-7.

20. Aray RJ, Hoff PM, De Castro G, *et al.* Ethical responsibility of phase 0 trials. *Clin Cancer Res* 2009 ; 15 (3) : 1121.

21. The Lancet. Editorial. Phase 0 trials: a platform for drug development? *Lancet* 2009 ; 374 (9685) : 176.

22. Rowan K. Oncology's first phase 0 trial. *J Natl Cancer Inst* 2009 ; 101 : 978-9.

23. LoRusso MP. Phase 0 clinical trials: An answer to drug development stagnation? *J Clin Oncol* 2009 ; 27 : 2586-8.

24. Kummar S, Doroshow JH, Tomaszewski JE, *et al.* Phase 0 clinical trials: Recommandations from the task force on methodology for the development of innovative cancer therapies. *EJC* 2009 ; 45 : 741-6.

Partie VI
Aspects pratiques

Gestion des données

L. Roca, C. Berneur-Morisseau

Une gestion optimale des données doit garantir une restitution totale et fiable de ces données au promoteur en respectant les bonnes pratiques cliniques et la réglementation en vigueur. Elle doit également assurer la sécurité des données et leur conservation sur support informatique.

La qualité des données rendues au promoteur est primordiale pour garantir une analyse fiable, permettant de répondre avec efficience à la question posée par l'étude.

Définitions

Le data-management

Le data-management comprend l'ensemble des actions de gestion et de traitement des données contribuant à assurer la documentation et la qualité de la base de données (BDD) cliniques dans laquelle seront reportées les informations recueillies au cours des essais thérapeutiques.

Pour les études régies par la Loi de Santé publique du 9 août 2004, le data-management doit :
- garantir la confidentialité des données médicales gérées et transmises ;
- garantir la qualité des données recueillies selon les règles des bonnes pratiques cliniques ;
- assurer la traçabilité de toutes les interventions sur la base de données.

Le data-management en recherche clinique n'est pas une étape isolée dans la gestion d'un essai, mais il s'intègre dans un processus multidisciplinaire, incluant de fortes collaborations en amont (avec les investigateurs, les responsables de projet, les attachés de recherche clinique (ARC), les opérateurs de saisie) et en aval (avec les statisticiens). Les différentes étapes de l'activité sont encadrées par des **procédures opératoires standards** (POS) pour assurer la qualité du traitement des données. La traçabilité tout au long du processus data-management est primordiale et inclut de nombreux modes opératoires, documents de suivi et documents de contrôle de la qualité.

Le data-manager

Le data-manager en recherche clinique est chargé de concevoir et de mettre en place les structures des bases de données pour les essais cliniques. Il veille également à gérer ces bases de données et

en assure la sécurité et la confidentialité. Le travail du data-manager passe par l'acquisition, la gestion, le codage, la validation, l'extraction et le stockage des données cliniques sans oublier la production de documents réglementaires relatifs à l'étude. Les points majeurs du métier sont :
• la relecture du protocole ;
• la participation à la rédaction du cahier d'observation (CRF pour *Case Report Form*) ;
• la rédaction du plan de data-management (PDM) ;
• la mise en place de la structure informatique de la base de données et définir sa gestion :
– annoter le CRF,
– construire la base de données,
– définir les modalités de la saisie et les règles de codage des données dans la base,
– maintenir et mettre à jour la structure de la base et sa documentation en conformité avec les évolutions du protocole,
– rédiger le plan de validation des données en parallèle *(Data Validation Plan)* et créer des programmes de contrôle de présence, condition et cohérence permettant de détecter les anomalies dans la base qui en découlent, valider et éditer les demandes de clarifications (DCF pour *Data Clarification Form*),
– organiser/contrôler le circuit des CRF et des DCF,
– participer à la revue des données en aveugle *(blind review)*,
– geler la base,
– transmettre une copie de la base au statisticien pour les analyses,
– archiver *les données*, les programmes de gestion et de contrôles informatiques,
– réaliser un contrôle qualité de toutes les étapes de la chaîne : création de la structure, saisie des données, création des contrôles de cohérence ;
• la rédaction les procédures et modes opératoires spécifiques de gestion des données et les tester :
– tenir à jour le cahier de codage/carnet de validation,
– mettre à disposition des bilans d'avancement de l'étude,
– assurer l'archivage des documents papier de l'étude concernant le data-management,
– assurer la traçabilité des changements éventuels de logiciels.

Le plan de data-management

Le plan de data-management (PDM) est une documentation obligatoire dans la gestion des données d'un essai clinique. Ce document regroupe les informations nécessaires à la gestion des données d'une étude et planifie les actions relatives à la gestion des données. Écrit par le data-manager de l'essai, le PDM est ensuite validé par le statisticien et l'ARC avant d'être transmis au promoteur de l'étude, faisant lieu de document repère entre les différentes parties. Les chapitres principaux sont les suivants :
• *synopsis de l'étude :* titre de l'essai, nombre de sujets inclus, critère principal, analyse intermédiaire, s'il y a lieu, et analyse finale ;
• *versionning :* dates et raisons des évolutions du document ;

- *interlocuteurs :* informations relatives aux personnes impliquées en termes de gestion des données dans l'essai (noms, e-mail, téléphones du promoteur, coordinateur, data-manager, statisticien et responsable du monitoring) ;
- *logigramme :* tableau indiquant sous forme d'étapes séquentielles et chronologiques les actions relatives à la gestion des données qui seront réalisées par les différents participants ;
- *définition des outils :* les logiciels utilisés et modalité d'acquisition des données ;
- *annotation du CRF :* description des champs, des variables calculées ou dérivées, listes de codage ;
- conventions générales de saisie : définition des règles de saisie adoptées pour l'étude et de remplissage des CRF ;
- *définitions des circuits de documents :* CRF, DCF, *tracking*, définition de la fréquence d'envoi de documents clefs au promoteur (états d'avancement) ;
- *validation des données :* plan de validation des données (tests de cohérences, listings) ;
- *rapport final de data-management :* compte rendu des actions lors du data-management (déviations, nombre et type de DCF, listes) ;
- *fin d'étude :* gels, dégels prévisionnels, archivage ;
- *conventions de transfert :* nature, modalité, fréquence et destinataire de transfert ;
- *documents associés :* plan de validation des données, liste des corrections auto-évidentes (SEC pour *Self Evident Corrections*), CRF annoté, guides ;
- *contrôle qualité :* définition des modalités de l'ensemble des contrôles [de la structure, de la saisie des données, historique des opérations effectuées sur la base (audit-trail), etc.].

Recueil des données

Le CRF est le document papier ou informatisé (e-CRF pour *electronic Case Report Form*) contenant l'ensemble des données recueillies tout au long de l'essai en adéquation avec les informations spécifiées dans le protocole. En termes d'exigences réglementaires, le CRF permet de conserver la trace authentifiée du recueil de ces données par l'investigateur, son contenu ne se substitue pas au dossier source.

Ces données ainsi recueillies sont ensuite utilisées pour l'analyse décrite dans le protocole. La qualité du CRF est donc primordiale pour répondre correctement aux objectifs de l'étude.

Élaboration

Le CRF est le résultat d'un travail collaboratif entre le coordonnateur, l'ARC, le data-manager et le statisticien. La version du CRF doit être clairement identifiée. L'identification de l'étude et du patient doit être présente à chaque page. Il est conseillé de recueillir les données non calculées ou non issues d'une transformation (par ex. date de naissance plutôt que l'âge).

Afin de limiter les erreurs de remplissage, quelques règles d'élaboration sont à respecter :
- les aide-mémoires et les modalités de remplissage doivent être écrits hors des fiches en intercalaire ;

- les questions doivent être claires (courtes, éviter les mots ambigus et les abréviations, ne poser qu'une question à la fois, ne pas suggérer la réponse dans la question, préférer des réponses fermées ou des Oui/Non dans les questionnaires) ;
- le CRF ne doit pas comporter le relevé de données non prévues par le protocole.

Format

Le CRF est créé au minimum sur du papier dupliqué (un exemplaire des fiches est conservé dans le cahier par l'investigateur, l'autre exemplaire est envoyé à l'équipe responsable du traitement des données. Il faut homogénéiser autant que possible la formulation et la présentation des cahiers d'observation. Le format défini est utilisé tout au long du CRF de manière constante.

- *Les variables qualitatives :* en questions fermées avec liste de réponses prédéfinies. Éviter le précodage numérique à recopier qui est source d'erreur mais faire figurer les codes associés aux libellés dans des cases à cocher.

- *Les variables quantitatives :* préciser l'unité, le format (nb de décimales). Utiliser les peignes.

- *Les questions ouvertes :* ce type de question nécessite un codage ultérieur pour pouvoir être analysé (à limiter au maximum).

- *Les dates :* utiliser les peignes au format défini, par exemple en France : JJ/MM/AAAA.

- Afin de minimiser les erreurs de saisie, les CRF doivent contenir les précodages nécessaires à la saisie : ex : (1) homme, (2) femme.

Modification du CRF en cours d'étude

En cas d'ajout d'item ou de modalité, un nouveau numéro de version du cahier ou de la fiche doit être édité, et une nouvelle soumission aux autorités compétentes doit être faite. Ce changement engendre aussi des modifications de la base de données et des contrôles qui en découlent.

Obligations réglementaires

- Anonymisation : les données recueillies ne doivent pas être nominatives.

- Signature de l'investigateur : certaines fiches du CRF (randomisation, clôture, examens complexes) doivent être datées et signées par l'investigateur.

Spécificité du e-CRF

Un e-CRF est une interface permettant le recueil, le contrôle et l'exploitation des données d'une étude clinique à distance *via* le réseau Internet, selon un protocole d'accès et de transfert sécurisé permettant ainsi de garantir la confidentialité (accès limité aux personnes autorisées), l'intégrité (données complètes et authentiques) et l'accessibilité (données disponibles aux personnes habilitées).

En fonction des ressources humaines, matérielles, le logiciel et les bases de données peuvent être hébergés par un prestataire de service externe ou en local.

En termes de bonnes pratiques cliniques, l'investigateur doit conserver une trace papier signée des e-CRF. L'e-CRF doit donc pouvoir être imprimé et signé pour répondre à cette recommandation.

Annotation du CRF

L'annotation du CRF permet de mettre en correspondance une variable du CRF avec un champ/ item de la base de données. L'annotation identifie les métadatas *(littéralement « données à propos d'une donnée »)* relatives à chaque variable et à leur organisation de stockage (par ex. *les métadonnées incluent le nom de chaque table*). Ces métadonnées sont un élément essentiel de l'architecture de la base *(figure 1)*.

L'annotation du CRF est une étape importante car elle détermine notamment les étapes de design et d'export de la base. La mise en place de standard pour le CRF et pour son annotation permet de simplifier le processus de création et de réalisation de la base de données.

Élaboration de la base de données

Les systèmes de gestion de bases de données sur lesquels reposent les interfaces sont généralement ORACLE, MYSQL, SQL SERVER.

Avant de saisir des données cliniques dans la base, la déclaration du traitement des données de l'essai doit avoir été faite (par le promoteur) auprès de la Commission nationale de l'informatique et des libertés (CNIL) ; il existe une procédure de référence pour faire cette déclaration avec un engagement des signataires.

IE

IEORRES *(OUINON)*

N° Item	CRITERES D'INCLUSION IECAT *(IECAT)*	Oui	Non
1 IESEQ	IETEST	❑1	❑0
2		❑1	❑0

IE : table IE (*Inclusion Exclusion criteria*) ; IESEQ-IECAT-IETEST-IEORRES : noms de variables ; *IECAT-OUINON* : bibliothèque de codage associée à la variable.

Figure 1. Architecture de base de données.

Pour la traçabilité, l'audit-trail automatique du logiciel de gestion de BDD permet d'enregistrer la totalité des opérations effectuées sur la base.

La sécurité de l'accès à la base doit être assurée (identification, mot de passe, cryptage des données, déconnexion automatique en cas d'inutilisation prolongée, etc.) afin de répondre aux exigences réglementaires : recommandations de la *Food and Drug Administration* (FDA) concernant les systèmes informatisés pour la gestion des essais cliniques *(Guidance for Computerized systems Used in Clinical Trials)* et signature électronique (21 CFR part 11) et aux normes internationales (CDISC, ICH, BPC).

Conception de la base de données

Le design consiste à mettre en place une interface d'acquisition des données grâce à un masque de saisie, dans le but de stocker les données sources d'une étude sur un support informatique. Ce design est réalisé à partir du CRF annoté. Il existe deux systèmes d'organisation des données dans les tables :

• Le premier dit **normalisé** contient peu de colonnes et beaucoup de lignes. Cette structure implique des récurrences par patient. Ce format est le plus utilisé par les standards CDISC :

Patient	Examens	Résultat	Unité
001	poids	60	kg
001	taille	168	cm

• À l'inverse, la version **non normalisée** est caractérisée par une seule ligne par patient, c'est-à-dire peu de ligne et beaucoup de colonnes :

Patient	Poids	Unité du poids	Taille	Unité de la taille
001	60	kg	168	cm

Validation de la base de données

La validation de la structure de la base de données est réalisée le plus souvent à partir de patients tests, listes de champs.

Acquisition des données

La saisie des données se fait par l'interface du masque de saisie (écran BDD ; page web) ou par import direct dans la base.

Saisie des données

La saisie peut être faite de plusieurs façons :
• soit par l'investigateur du centre à partir d'écran d'e-CRF : on parle alors de **saisie en direct**. Cette saisie se fait à distance ;
• soit par des professionnels de la saisie informatique (opérateurs de saisie) à partir de CRF papier : on parle alors de **saisie différée**.

Dans ce dernier cas, la saisie est centralisée et peut être réalisée :
• soit en *simple saisie* par un seul opérateur. En termes de data-management, les contrôles doivent être plus nombreux et un audit plus vaste des données est nécessaire ;
• soit en *double saisie* par deux opérateurs différents :
– double saisie indépendante : la saisie est faite en aveugle par deux opérateurs indépendants. Les erreurs sont traitées par le data-manager après mise en évidence par un programme de comparaison des données,
– double saisie interactive : les discordances détectées sont corrigées interactivement à la seconde saisie. C'est donc le second opérateur qui est arrêté dans sa saisie en cas de saisie différente de la première. En cas de non-résolution de la différence, une demande de clarification (*query*) est émise par le data-manager.

Il est avéré qu'une double saisie reste le garant d'une bonne qualité de la saisie des données avec un taux d'erreur entre 0,1 et 0,2 %. Bien que la saisie soit plus longue, les étapes suivantes du data-management sont accélérées.

Afin d'améliorer la qualité et la reproductibilité de la saisie, un document recensant les conventions de saisie et les spécificités de l'étude est rédigé.

La saisie doit être le reflet du document original et aucune interprétation ou correction ne doit être faite. Seules les modifications autorisées par les « conventions de saisie » ou par les « conventions avant gel de base » ou les « SEC » *(Self Evident Corrections)* donnent lieu à une donnée différente entre la base et le CRF. Un audit-trail de la saisie doit être disponible pour la traçabilité des opérations effectuées.

Import direct des données

Les données peuvent être importées directement dans la BDD grâce à un procédé permettant de convertir les formats d'une base de données vers une autre (processus de *mapping*). Une attention particulière sera portée au contrôle qualité de ces données importées.

Gestion de la base de données

Afin de s'assurer de la présence et de la cohérence des données dans la base, de nombreux contrôles et requêtes sont définis. Chaque incohérence induit une demande de clarification *(query)* sur les données incertaines. Un bordereau regroupant les demandes d'un patient (DCF pour *Discrepancy Clarification Form*) est alors adressé à l'investigateur concerné.

Identification des tests/contrôles

Ces tests sont définis au début de l'étude lors la structuration de la base de données par le data-manager et validés par l'équipe clinique. Ces contrôles sont regroupés dans un « carnet de test ». Chaque test est clairement identifié (page de CRF, table, n^o de test) et un message demandant une correction lui est attribué. Le test est écrit de façon compréhensible (dans un langage non programmé) et les messages adressés ne doivent pas guider la réponse.

Contrôle automatique

La majorité des tests peuvent être programmés et édités de façon automatique. Ces contrôles doivent concerner *a minima* les données suivantes : critères d'inclusion/non-inclusion, délais, paramètre principal d'évaluation, traitements administrés, données manquantes non confirmées par l'investigateur.

La validation des programmes de test doit être faite par un data-manager à l'aide de jeux de données et/ou sur un nombre de cahiers établi.

L'ensemble des DCF éditées doit être revu/validé par le data-manager avant envoi aux centres investigateurs.

***Queries* résolues en interne :** certaines *queries* peuvent être résolues directement en interne par des corrections dites « évidentes » (SEC) *(figure 2)*. La liste des SEC spécifiques à l'étude est établie au moment de la rédaction du carnet de test et envoyée à l'investigateur pour validation et signature.

Revue manuelle

Certaines anomalies présentes dans la base de données ne peuvent pas être mises en évidence au moyen d'un test logique programmé et nécessitent l'envoi de DCF réalisées manuellement. Le statisticien ou le data-manager programme alors par un système externe ou non au logiciel de data-management des analyses simples ou des listings de revues de données pour mettre en évidence ces incohérences. Ces DCF, comme les automatiques, doivent être enregistrées dans le système. Elles sont ensuite traitées suivant la même logique que les DCF automatiques.

Examen clinique réalisé .... |__| 0 = non, 1 = oui **champ 1 vide**
Si oui, poids (kg) .... |56| **champ 2 associé rempli ⇒SEC sur champ 1**
Correction SEC :
Examen clinique réalisé .... |1| 0 = non, 1 = oui **champ 1 corrigé par SEC**

Figure 2. Exemple de SEC.
*Examen réalisé est « **vide** »* alors que le résultat associé est renseigné. Une SEC permettra de corriger le champ Examen réalisé en *« **oui** »*.

Gestion des DCF

L'envoi de DCF est tracé par un n° de DCF/bordereau et une date d'envoi.

La réponse aux DCF peut se faire par fax, papier ou e-mail selon les méthodes de chacun. La réception est tracée et enregistrée. Après saisie, la DCF originale est archivée par le promoteur. L'investigateur en conserve une copie.

L'état des DCF doit pouvoir être connu tout au long de l'étude, notamment :
- *état des corrections émises :* identifiée ; traitée et toujours en cours ; envoyée sur site ; duplicata ; réceptionnée (mais non encore traitée) ;
- *type de résolution :* résolue (avec un bordereau) ; non résolue (mais incorrecte) ; SEC ;
- *source de résolution :* investigateur ; data-manager.

Pour une bonne gestion, les données doivent être contrôlées le plus tôt possible dans l'étude pour permettre une identification rapide des problèmes liés au format du CRF ou au remplissage de celui-ci. Il est important aussi, en fin d'étude, de revoir les mesures de corrections qui permettent d'améliorer les études suivantes, les SEC, la programmation des tests de cohérence : identifier les tests le plus souvent émis ou les champs le plus souvent liés à une erreur ; le pourcentage de *queries* résolues en interne et le pourcentage de *queries* émises.

Correction de la base de données

Comme le processus de saisie, la correction de la base doit se faire selon une procédure précise afin d'assurer la traçabilité.

L'ensemble des opérations effectuées sur la base (connexion, modification) doit être enregistré et consultable à tout moment. L'ancienne valeur, la nouvelle valeur, la date de modification et la personne qui a effectué la modification, ainsi que la raison de la modification doivent apparaître (audit-trail).

Après chaque modification, les contrôles de cohérence doivent être refaits sur les nouvelles données, jusqu'à ce que la totalité des incohérences de la base soient résolues. On parle dans ce cas d'une base « propre ».

Contrôle qualité

L'assurance qualité est un processus qui est basé sur des standards, des procédures permettant d'atteindre et de maîtriser un niveau de qualité des données souhaité. Les contrôles de qualité s'intègrent à cette action. L'assurance qualité est documentée dans le PDM.

Audit de la base de données

Le contrôle qualité des données est représenté par un contrôle des données de la base par rapport aux données sources (CRF + *queries*). Cet audit de base est défini dans un plan d'audit (contenant la taille de l'échantillon d'audit, la définition d'un taux acceptable d'erreurs et une solution face à un taux trop important d'erreur), et un résumé doit être documenté dans un rapport d'audit (contenant le nombre total d'erreurs, le taux d'erreur et les actions prises) et doit être réalisé avant la fin de la saisie. En cas de mauvais résultats et après correction des variables auditées, un second échantillon est défini où le nombre de champs peut être augmenté. La nature des erreurs est identifiée pour définir des actions correctives.

Par exemple : L'échantillon pour un audit est généralement de 10 % de l'ensemble des données (contrôle qualité horizontal) et 100 % des données du critère principal (contrôle qualité vertical). Le taux d'erreur est généralement considéré comme acceptable entre 0,1 et 0,5 % pour le contrôle horizontal.

Audit-trail

L'audit-trail proposé par les logiciels de data-management trace toutes les opérations effectuées sur la base de données afin de repérer les connexions, les modifications et leurs raisons pour chacune des variables.

Tracking des pages de CRF et DCF

L'étape de *tracking* correspond au contrôle de la réception de documents ; les informations sont tracées et reportées dans un tableau de bord qui peut être un fichier électronique, suffisant pour indiquer qu'un *tracking* des pages de CRF ou DCF est mis en place (on peut s'attendre à avoir des rapports édités et signés à la fin de l'étude ou encore des documents donnant l'état des corrections et du statut des pages).

Fin d'étude

Revue finale des données *(blind review)*

Il s'agit de passer en revue, avant le gel de la base, l'ensemble des sujets/thèmes ayant trait à la gestion des données avant l'analyse statistique.

Objectifs

Faire un bilan final de la totalité des données de l'étude (sans avoir connaissance des codes de traitement de chacun des sujets si l'essai est randomisé afin de garantir indépendance et objectivité) afin d'identifier les problèmes restant et statuer sur chacun d'entre eux.

Les points abordés

- Rappel du protocole : objectif principal, amendements, attribution des unités de traitement.

- Mode de gestion des données : la structure de la base de données, les règles de saisie, les règles de validation des données et les corrections automatiques, les documents de contrôle de qualité, le codage, les données transférées et calculées, la validation, les déviations, les bornes biologiques.

- Bilan des données de l'étude : nombre de centres, nombre et statut des patients, observance, durée d'exposition au traitement, paramètres d'efficacité et de sécurité, effets indésirables, point sur les DCF, déviations, données manquantes, nombre de patients pour les analyses en intention de traiter et per protocole.

Déroulement

L'ensemble des acteurs de l'étude concernés sont présents à cette revue. Des listings décrivant les données importantes sont édités par le data-manager et/ou le statisticien. Les décisions qui seront prises sur les données doivent être validées par chacun des participants et définies avant la revue des données (conventions précisées dans le PDM). Les éventuelles modifications apportées seront intégrées dans la base et documentées dans le rapport final de data-management.

Gel de la base de données

Avant de pouvoir geler une base, de nombreuses tâches sont à effectuer :

- l'ensemble des données doit être récupéré, validé et signé ;
- les DCF doivent être résolues en totalité ;
- les données issues de la base de pharmacovigilance doivent être réconciliées avec celles issue de la base de données le cas échéant.

Une fois ces actions réalisées, il est nécessaire de « verrouiller » l'accès en écriture aux tables contenant ces données. Le gel de la base consiste à supprimer les droits en écriture sur ces tables. L'étude sera alors en « lecture seule ». Le gel de la base doit se faire à la demande du promoteur.

Il est possible de « dégeler » une base, mais cette procédure doit rester exceptionnelle et être documentée avec précision.

Levée d'aveugle (le cas échéant)

Cette étape consiste à dévoiler le code de randomisation afin d'identifier le traitement reçu par le patient.

Sauvegarde et archivage des données

Les données archivées doivent pouvoir être consultées à tout moment au cours du temps sur demande (patient, autorités sanitaires).

En Europe, la durée d'archivage des documents des essais cliniques par le promoteur est de 15 ans minimum (40 ans pour les médicaments dérivés du sang). La documentation relative à la fabrication des différents lots de médicaments doit être conservée, selon l'Union européenne, au moins 2 ans après la fin de l'étude. Une procédure d'archivage doit être mise en place selon les possibilités de chacun (*a minima* définition du classement des documents de l'étude ainsi que les modalités d'archivage) et testée régulièrement pour assurer une restauration rapide et intègre des données.

Il est conseillé d'archiver au moins 2 exemplaires de la base qui seront stockés sur des supports différents (disques, magnétiques ou optiques) et dans des endroits différents. Par mesure de précaution, il est recommandé de conserver une copie du logiciel utilisé pour le traitement des données. Il faut donc prévoir une sauvegarde des données et du système.

Sécurité des données

La protection physique est aussi importante que la protection logique pour assurer l'intégrité et la confidentialité des données.

Sécurité physique

La sécurité physique est la sécurité des centres de traitement de l'information. Elle concerne les protections contre les intrusions, le vol, les accidents, les risques naturels.

Elle peut conduire à l'utilisation de clés, badges, cartes d'accès, gardes, détecteurs.

Sécurité logique

La sécurité logique est la sécurité fournie par le système d'exploitation de la machine et les logiciels de base. Elle est sous la responsabilité de l'administrateur système.

Il s'agit de pouvoir assurer les contrôles d'accès aux données (mot de passe, protocole d'accès ou mode de transmission sécurisés).

Pour en savoir plus

Livres

- Prokscha S (ed). *Practical Guide to Clinical Data Management*. 2nd edition. Boca Raton : CRC Press/Taylor & Francis, 2007, 238 pages.
- Rondel RK, Varley SA, Webb CF (eds). *Clinical Data Management*. 2nd edition. West Sussex : John Wiley & Sons, 2000, 354 pages.
- Spriet A, Dupin-Spriet T (eds). *Bonne pratique des essais cliniques des médicaments*. Bâle : Karger, 2004, 273 pages.

Textes réglementaires

- 21 CFR part. 11 Federal Register : March 20, 1997 ; Vol. 62 – N° 54 : 13430-13466.
- Commission nationale de l'informatique et des libertés. *Méthodologie de référence MR-001 pour les traitements de données personnelles opérés dans le cadre des recherches biomédicales*. CNIL, octobre 2010.
- *International Conference on Harmonisation of technical requirements for registration of pharmaceuticals for human use* :

– ICH topic E6(R1). Note for Guidance on Good Clinical Practice. July 1996 ;

– ICH documents : M1 MedDRA.

- *Efficacy guidelines : Good Clinical Practice : E2 - Clinical Safety Data Management* :

– E3 – Structure and Content of Clinical Study Reports ;

– E6 – Good Clinical Practice ;

– E7 – Studies in Support of Special Populations/Geriatrics ;

– E8 – General Consideration of Clinical Trials ;

– E9 – Statistical Principles for Clinical Trials.

- Loi n° 2004-806 du 9 août 2004 relative à la politique de santé publique.
- Directive 2001/20/CE article 1.2.
- Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés.

Organismes et dictionnaires

- CDISC : Clinical Data Interchange Standards Consortium (www.cdisc.org).
- CNIL : Commission nationale de l'informatique et des libertés (www.cnil.fr).
- Dictionnaires de toxicités : OMS, NCI, RTOG, CTC.
- Dictionnaires de maladies : CIM9, CIM10, ECOG, SIOP.

Modalités de randomisation

C. Ferlay, S. Gourgou-Bourgade

Ce chapitre présente le principe et les modalités du tirage au sort dans les essais cliniques.

Le principe général de l'essai thérapeutique contrôlé est d'évaluer l'efficacité d'un traitement en comparant des groupes de patients ayant reçu des traitements différents mais comparables par ailleurs.

Afin de mettre en évidence cet effet propre au traitement, il est nécessaire de comparer le groupe de patients recevant ce traitement à un groupe comparable en tout point, excepté pour le traitement administré. L'attribution du traitement « au hasard » *(random)* s'est imposée comme le seul moyen d'assurer la comparabilité initiale entre les groupes. On parle de randomisation ou tirage au sort ou allocation aléatoire.

Malgré cela, il peut arriver que le seul fait du hasard produise des différences notables pour des caractères importants, mais on pourra en tenir compte au moment de l'analyse (notion d'ajustement).

Un tirage au sort plutôt qu'une étude observationnelle

Afin de comparer l'efficacité d'un nouveau traitement par rapport à un traitement de référence, une possibilité serait de comparer un groupe de patients recevant le nouveau traitement et, par exemple, les patients qui durant l'année précédant l'introduction du nouveau médicament ont reçu l'ancien traitement. Ce groupe contrôle, appelé **contrôle historique**, ne remplit pas les conditions pour être un comparateur acceptable. En effet, rien ne garantit que les anciens patients soient comparables aux nouveaux : la prise en charge des patients a pu évoluer au cours du temps. De plus, les autres traitements concomitants ont certainement eux aussi évolué et donc un meilleur résultat obtenu sur les nouveaux patients signifie peut-être tout simplement que la prise en charge des patients s'est améliorée, sans que le nouveau traitement soit meilleur que l'ancien.

Un groupe contrôle doit être constitué de patients contemporains aux patients inclus dans le groupe traité.

Une autre possibilité serait de comparer les patients traités par le traitement standard dans un service hospitalier A et les patients du service B traités par le nouveau médicament. Ce procédé est inacceptable car rien ne garantit que les patients recrutés dans ces deux services soient similaires et pris en charge de la même manière.

Différents moyens qui ne sont pas basés sur un tirage au sort ont parfois été utilisés pour attribuer un traitement aux patients ; en voici quelques exemples :
• l'alternance : administrer chacun des traitements alternativement à un malade sur deux ou administrer le traitement A tel jour et B le lendemain et ainsi de suite ;
• l'utilisation de l'année de naissance : donner par exemple le traitement A aux sujets nés les années paires et le traitement B aux sujets nés les années impaires.

Dans ces deux cas, le médecin connaît, à l'avance, le traitement à administrer au malade : cela peut influencer sa décision. En effet, en fonction de certains *a priori* sur les traitements comparés, le médecin peut, dans le 1er cas, choisir de retarder l'inclusion du patient dans l'essai ou, dans le 2e cas, décider de ne pas inclure le patient dans l'étude. Ces situations sont en contradiction avec la **clause d'ignorance** qui veut que le médecin ignore le traitement que recevra son patient lorsqu'il décide de l'inclure ou non dans un essai.

Tous ces exemples mènent à la conclusion que seule l'attribution au hasard permet :
– de répartir de manière équilibrée les caractéristiques des patients (comme l'âge, le sexe, etc.) susceptibles de produire des biais dans les résultats de l'analyse statistique ;
– d'éviter les biais de sélection attribuables aux participants (par rapport à leur opinion sur les traitements) ;
– et ainsi donc de se mettre dans la situation la plus favorable pour conclure sur le lien causal entre le traitement et l'effet mesuré.

Aveugle ou insu

Afin de maintenir la comparabilité des groupes tout au long de l'essai et garantir l'égalité dans l'appréciation ou l'évolution de la maladie, l'attribution du traitement doit se faire sans que le patient et/ou le médecin en connaissent la nature **(aveugle ou insu)**. Dans le cas contraire, on dit que l'essai est réalisé en ouvert.

Simple insu : la personne qui se prête à la recherche n'est pas informée de la nature du traitement qui lui est attribué.
Double insu : ni la personne qui se prête à la recherche, ni le médecin, ni le moniteur, ni même parfois la personne qui analyse les données ne sont informés de la nature des traitements attribués.

Cette méthodologie est généralement utilisée dans les essais contre placebo mais est rarement applicable dans les essais d'évaluation de traitements dont les effets ou la nature du traitement sont clairement identifiables (cytotoxique ou cytostatique à effets secondaires spécifiques, acte chirurgical, radiothérapie).

La création de listes de randomisation

Afin d'attribuer au hasard un traitement, il s'agit d'établir une liste d'allocation des traitements de l'essai. Cette liste sera utilisée pour répartir les patients dans chacun des groupes à comparer. Dans le cas des études ouvertes, la liste contiendra un numéro de patient (attribué lors de l'inclusion) et la nature du traitement correspondant. Pour les études en aveugle, la liste sera composée d'un numéro de patient auquel sera associé un numéro de traitement – il est conseillé de ne pas utiliser le n° de patient comme n° de traitement – *(figure 1)*. Des listes de conditionnement (attribution d'un n° de traitement) ainsi qu'une liste complémentaire (liste de levée d'insu) mettant en relation le numéro de traitement et la nature du traitement seront également éditées. La liste de levée d'insu sera tenue secrète et détenue uniquement par le centre de levée d'insu.

Avant d'établir la liste d'allocation, selon le type d'essai, il est nécessaire de définir certains paramètres, notamment la présence de variables de stratification correspondant aux facteurs pronostiques connus et l'utilisation de blocs de permutation afin de maintenir la clause d'ignorance.

Stratification

La randomisation assure qu'en moyenne les patients sont comparables. Cela n'assure pas forcément que critère par critère la répartition entre les bras est identique. Parfois, il arrive malgré le tirage au sort que les groupes de patients ne soient pas semblables, l'interprétation des résultats devient alors compliquée. Par exemple, il est arrivé qu'à la fin d'un essai comparatif, les résultats de survie soient similaires et que pourtant *a posteriori* les groupes n'étaient pas comparables pour un critère initial important (nombre de patients avec des métastases ganglionnaires plus élevé dans le bras expérimental). De ce fait, l'absence de différence en survie est-elle liée à l'effet propre du traitement ou au fait que les patients avaient des métastases ganglionnaires à l'inclusion ?

Étude en ouvert	
N° du patient	Nature
1	A
2	B
3	B
4	A
5	B
6	B
7	A
8	A

Étude en aveugle		
N° du patient	N° du traitement	Nature
1	2569	A
2	2148	B
3	4587	B
4	7896	A
5	6973	B
6	1236	A
7	7893	B
8	4569	A

↳Liste de correspondance secrète

Figure 1. Exemples de listes de randomisation.

Afin de limiter ces cas, dans un essai randomisé, s'il existe un facteur pronostique reconnu (par ex. métastases ganglionnaires : oui *vs* non), il est préférable de le définir comme une variable de stratification. Ainsi, deux listes d'allocation seront générées, une liste pour les patients ayant des métastases ganglionnaires et une liste pour les patients n'en ayant pas. Cette façon de procéder assure qu'à la fin des inclusions la proportion de patients ayant des métastases ganglionnaires sera similaire dans chacun des bras de traitement.

En cas de variable pronostique ou de facteur confondant connu comme l'effet centre – dans les essais thérapeutiques, le centre investigateur est souvent une variable de stratification –, la stratification assure que la répartition entre les bras sera respectée sur le critère de stratification. Elle augmente ainsi la comparabilité des groupes et la puissance statistique de l'étude en constituant des groupes plus homogènes. En pratique, le recours à la stratification doit être justifié et ne pas être excessif. L'utilisation de plus de deux ou trois facteurs de stratification est rarement nécessaire et justifiée. En effet, au-delà de trois niveaux de stratification, la mise en pratique devient complexe car cela multiplie les listes de randomisation malgré l'utilisation de logiciels. En outre, il se peut que certaines combinaisons de catégories des facteurs de stratification soient mal représentées, ce qui peut entraîner un déséquilibre dans le nombre de patients entre les deux groupes de traitement.

La présence ou non de variable de stratification va déterminer le nombre de listes de randomisation à générer (une liste par combinaison de strate). Par exemple, pour un essai randomisé auquel trois centres participent et pour lequel la variable sexe (deux modalités) est un facteur de stratification, six listes distinctes seront générées par tirage au sort :
– liste 1 : patients inclus dans le centre 1, de sexe masculin ;
– liste 2 : patients inclus dans le centre 1, de sexe féminin ;
– liste 3 : patients inclus dans le centre 2, de sexe masculin ;
– liste 4 : patients inclus dans le centre 2, de sexe féminin ;
– liste 5 : patients inclus dans le centre 3, de sexe masculin ;
– liste 6 : patients inclus dans le centre 3, de sexe féminin.

Plusieurs méthodes

Il existe plusieurs méthodes, dont les plus utilisées sont décrites ci-après, de la randomisation simple à la randomisation dynamique plus complexe en cas de plusieurs facteurs de stratification connus.

Les listes de nombres au hasard

On utilise les listes de nombre au hasard pour générer des listes de randomisation simple (sans notion de blocs). Elles donnent des suites de chiffres de 0 à 9, telle que chacun soit obtenu avec la même probabilité et de façon indépendante. En pratique, on définit au préalable, par exemple qu'un chiffre pair correspond au traitement A et qu'un chiffre impair correspond au traitement B. Ensuite, on part d'un chiffre de la table et on lit les chiffres de sa ligne (ou de sa colonne) et des

suivantes dans un sens ou dans l'autre. Par exemple, d'après la *figure 2*, la suite **154776603** correspond à la suite BBABBAAAB. Le premier patient randomisé recevra le traitement B, le second le traitement B, le troisième le traitement A et ainsi de suite.

L'utilisation de cette méthode permet d'assurer l'imprévisibilité mais pas l'équilibre des bras de traitement alloué.

Les listes de permutation au hasard

Les listes de permutation au hasard sont utilisées afin d'obtenir des blocs de traitement équilibrés. Elles sont issues de processus aléatoires complexes qui fournissent des séries de nombres équilibrés tous les x éléments (6, 9, 20...), chaque nombre étant unique à l'intérieur de la série considérée. Prenons le cas d'un essai thérapeutique randomisé comparatif à deux bras. Les patients du premier bras recevront le traitement A et les patients du second bras recevront le traitement B. On décide d'équilibrer les traitements tous les 4 patients en utilisant une table de permutation à 9 éléments *(figure 3)*. On constate que chaque colonne contient les chiffres de 1 à 9 dans un ordre aléatoire (permutation). Il suffit alors d'établir une correspondance préalable de type A = 1 ou 2 et B = 3 ou 4, de choisir une colonne et d'en déduire la séquence de traitements correspondante sans tenir compte des chiffres autres que 1, 2, 3 et 4. Ainsi la colonne **478621359** (2[e] colonne en **gras**) correspond à la séquence BAAB et ainsi de suite en lisant les colonnes suivantes. Au bout de 4 patients inclus, on a bien autant de patients randomisés dans le bras A que de patients randomisés dans le bras B.

19853	06933	69767	88842
282**15**	**47766**	**03**076	25940
68517	67954	16570	72433
59002	26619	02930	83677
92531	70313	24969	14458
74348	66239	32704	41018
96194	15831	08968	45321
39588	57825	36521	85188
18313	82950	12335	32398
68012	52485	55139	73430

Figure 2. Extrait d'une table de nombres au hasard.

84991	43379	**93**459	32561
37185	71913	**69**815	64492
55662	88437	**25**731	55255
62439	65621	**34**563	93344
93228	29248	**42**642	79726
48774	12594	**76**127	28673
21857	37862	**58**296	47137
19316	56155	**87**384	11918
76543	94786	**11**978	86889

sens de lecture

Figure 3. Extrait d'une table de permutations au hasard à 9 éléments.

Ces tables peuvent également être utilisées pour des randomisations déséquilibrées, comme par exemple si on veut avoir pour un patient recevant le traitement A, 2 patients recevant le traitement B. Pour cela, on décide de faire des blocs de 6 et d'associer au traitement A les chiffres 1 et 2 et au traitement B les chiffres 3, 4, 5 et 6. Les colonnes 962347581 et 178624593 (deux premiers éléments de la 3e colonne en **gras**) correspondent à la séquence de traitements BABBBA ABABBB. On constate que tous les 6 patients, le ratio 1:2 est bien respecté.

Blocs de permutation

Un tirage aléatoire de 100 natures de traitement 50 A et 50 B peut conduire à la liste suivante : AAAAA... BBBBBB. Cette liste pose différents problèmes. En effet, si les inclusions sont arrêtées prématurément il y aura un déséquilibre entre les bras. De plus, en cas de période de recrutement assez longue, il peut y avoir une différence de prise en charge entre le début et la fin des inclusions et donc une différence de prise en charge entre les patients recevant le traitement A et ceux recevant le traitement B. Une telle série fait penser à une étude avec un contrôle historique où on traite d'abord les patients avec le traitement A, puis les suivants avec le traitement B. Ce sont pour ces raisons que l'on privilégie les randomisations par bloc.

L'utilisation de blocs a pour but de limiter les déséquilibres. Le déséquilibre entre les groupes ne peut pas dépasser « 1/2 la taille du bloc » (pour des blocs de 4 avec 2 A et 2 B, le déséquilibre ne peut être supérieur à 2).

Reprenons l'exemple de l'étude en ouvert de la *figure 1* : en utilisant des blocs de 4, tous les 4 patients il y aura autant de traitements A attribués que de traitements B. Au pire des cas, si on inclut seulement 6 patients dans l'étude, il pourrait y avoir 2 patients dans le groupe traitement A et 4 patients dans le groupe traitement B, soit un déséquilibre de 2 patients (1/2 * 4).

Il est impératif que la taille des blocs ne soit pas divulguée aux investigateurs et donc pas notifiée dans le protocole. Cependant, un inconvénient de cette méthode très couramment utilisée est le risque d'apprentissage intuitif de la taille du bloc par les investigateurs et, par conséquent,

l'introduction toujours possible d'un biais dans l'allocation des traitements. Pour cette raison, il est vivement conseillé, surtout dans un essai en ouvert, de prendre des blocs de taille suffisamment importante ou de taille différente, de préférence aléatoire. La taille des blocs est également déterminée en fonction de l'effectif total de l'essai et pour les essais stratifiés par le nombre de listes à générer et du potentiel recrutement prévu par strate.

Cette méthode a des limites dans les essais de petite taille ou en cas d'analyse intermédiaire du fait de la taille des blocs qui peuvent mettre en cause l'imprévisibilité.

L'allocation dynamique

Cette méthode est souvent utilisée quand il y a plusieurs variables de stratification et par conséquent beaucoup de strates. L'utilisation de l'outil informatique permet de programmer des tirages au sort plus complexes comme le tirage au sort par **minimisation** [4, 6, 7] encore appelée randomisation adaptative ou dynamique. À chaque demande de randomisation, un programme informatique calcule en temps réel l'attribution du groupe qui garantit le meilleur équilibre possible entre les groupes. Pour le premier patient de l'essai, le traitement est alloué au hasard et ensuite à chaque patient supplémentaire, le traitement est alloué de manière à minimiser le déséquilibre entre les groupes en tenant compte des valeurs des variables de stratification du patient à randomiser et des patients déjà randomisés.

Exemple de randomisation par minimisation

Supposons un essai randomisé à deux bras (bras A *vs* bras B) stratifié selon quatre variables : âge (≤ 50 ans *vs* > 50 ans), sexe (homme *vs* femme), tabagisme (fumeur *vs* non-fumeur), stade de la maladie (I *vs* II *vs* III). Si en plus l'essai est multicentrique et donc stratifié aussi sur le centre, avec une randomisation classique il faudrait générer 24 listes par centre et donc il y aurait un risque important de déséquilibre si le potentiel d'inclusion par centre est faible même avec des blocs de petite taille. Pour simplifier, considérons cet essai comme monocentrique et supposons qu'après la randomisation de 30 patients, les effectifs soient ceux du *tableau I* et que le prochain patient à randomiser soit une femme de 52 ans fumeuse avec une maladie de stade I. Pour allouer le bras du nouveau patient, une des méthodes est de sommer les effectifs des 30 patients déjà randomisés qui ont les mêmes caractéristiques (sur les variables de stratification) que le nouveau patient. On obtient donc :

- bras A : 10 (sexe) + 8 (âge) + 10 (tabagisme) + 10 (stade de la maladie) = 38 ;
- bras B : 6 + 7 + 12 + 10 = 35.

Le déséquilibre est alors minimisé en allouant le nouveau patient dans le groupe avec le plus petit total (ou au hasard si les totaux sont les mêmes). Mais on peut également décider d'attribuer le patient au bras B (car c'est le plus petit total) avec une probabilité égale à 0,75 et de l'attribuer au bras A avec une probabilité égale à 0,25 pour laisser « au hasard » une part plus importante. C'est d'ailleurs cette option qui est recommandée (Pocock [4]).
Une fois le traitement alloué, les effectifs des variables de stratification sont mis à jour dans chaque bras et le processus est renouvelé à chaque nouveau patient.

Tableau I. Exemple pour une randomisation par minimisation : caractéristiques des 30 patients déjà randomisés.

		Bras A N = 15	Bras B N = 15
Sexe	Homme	5	9
	Femme	10	6
Âge	≤ 50 ans	7	8
	> 50 ans	8	7
Tabagisme	Fumeur	10	12
	Non-fumeur	5	3
Stade de la maladie	I	10	10
	II	4	5
	III	1	0

Il existe plusieurs méthodes de randomisation par minimisation [4, 7].

Une fois les paramètres de la randomisation définis (variables stratification, bloc, etc.), les listes de randomisations sont générées à partir de fonction prédéfinie dans les logiciels statistiques usuels, telle que « ALEA » sous Excel, « PROC PLAN » sous SAS, « RALLOC » sous Stata pour les techniques de randomisation par blocs.

Aujourd'hui, de plus en plus de logiciels de gestion de données cliniques ont des modules intégrés de randomisation utilisant ces principales méthodes. D'autres logiciels développés au niveau européen tels que TenAlea proposent aujourd'hui des modules intégrés pour la mise en œuvre de la randomisation, notamment par minimisation. Nous renvoyons le lecteur au chapitre consacré aux logiciels dédiés pour en savoir plus (chapitre VI.5 « Les logiciels », page 370).

Les recommandations

Pour la description de la randomisation dans le rapport des données des essais cliniques, l'énoncé CONSORT [8] préconise de préciser :
– la méthode utilisée pour générer la séquence aléatoire d'allocation des traitements, incluant le détail de la méthode (blocs, stratification) ;
– la méthode utilisée pour implémenter la randomisation (enveloppes, central téléphonique, fax, Internet) en précisant comment la séquence était dissimulée jusqu'à ce que l'intervention soit réalisée ;
– la personne qui génère le code d'allocation des traitements, qui inclut les patients et qui attribue le groupe de traitement.

L'énoncé préconise également de produire un diagramme (CONSORT Flow-Chart) résumant chaque étape d'inclusion/d'allocation/de suivi et d'analyse (*cf*. chapitre VI.4 « Plan statistique, rapport d'analyse statistique et rapport final d'essai », page 361).

En pratique : la réalisation de la randomisation d'un patient

À retenir

- La randomisation centralisée est indispensable pour assurer la clause d'ignorance.
- Le respect de la clause d'ignorance permet de ne pas sélectionner les patients (utilisation de l'aveugle si possible).
- La méthodologie doit être adaptée à chaque question scientifique et logistique (méthode, ratio, facteurs de stratification, faisabilité logistique).
- La randomisation garantit l'équilibre des bras à comparer afin d'obtenir un jugement de causalité.
- À ce jour, de nombreuses possibilités logicielles pour l'ensemble des méthodes proposées.
- La technique de minimisation est à recommander à partir de deux facteurs de stratification.

L'allocation centralisée est à privilégier dans le but de préserver le caractère imprévisible de la randomisation.

Après avoir vérifié l'éligibilité du patient, avoir expliqué les objectifs de l'étude et les traitements possibles, obtenu le consentement du patient, chaque investigateur contacte le centre de randomisation (par fax, téléphone, *via* Internet) qui lui communique à partir de la liste préétablie ou le programme informatique soit la nature, soit le numéro du traitement. En procédant de cette manière, seul le centre d'allocation détient la liste de randomisation et les investigateurs découvrent le traitement après inclusion et vérification des critères d'inclusion et d'exclusion du patient. De plus, le centre d'allocation est informé en temps réel des randomisations dans l'essai.

La randomisation d'un patient dans un essai doit avoir lieu le plus près possible du début du traitement afin d'éviter qu'un événement intercurrent (décès du patient ou prise de médicaments incompatibles) survienne entre le moment où son traitement a été alloué et la mise en route de ce dernier. Ces faits pourraient conduire à prendre en compte, dans l'analyse en intention de traiter, des patients ne recevant pas le traitement de l'étude.

Les pièges à éviter

- Divulgation de la taille des blocs.
- L'utilisation de facteurs de stratification est à recommander avec un nombre de facteurs pertinents raisonnable (techniques de minimisation).
- L'ajustement sur le centre ne sera pas possible à l'analyse, penser à stratifier sur ce facteur si nécessaire.

Références

1. Schwartz D, Flammant R, Lellouch J. *L'essai thérapeutique chez l'homme.* 2e édition. Paris : Flammarion Médecine-Sciences, 1992 : 71-8.

2. Laplanche A, Com-Nougué C, Flamant R. *Méthodes statistiques appliquées à la recherche clinique.* Paris : Flammarion Médecine-Sciences, 1987 : 6-11.

3. Huguier M, Flahault A. *Biostatistiques au quotidien.* Paris : Elsevier, 2000 : 127-33.

4. Pocock SJ, Simon R. Sequentiel treatment assignment with balancing for pronostic factors in the controlled clinical trial. *Biometrics* 1975 ; 3 : 103-15.

5. Gambotti L, Perol D, Chauvin F. Randomisation et tirage au sort. *La revue du praticien* 2004 ; Tome 18 : 648-9.

6. Douglas G Altman, J Martin Bland. Treatment allocation by minimisation. *BMJ* 2005 ; 330 : 843.

7. Signorini DF, Leung O, Simes RJ, Beller E, Gebski VJ. Dynamic balanced randomization for clinical trials. *Stat Med* 1993 ; 12 : 2343-50.

8. Begg C, Cho M, Eastwood S, *et al.* Improving the Quality of Reporting RCTs. *JAMA* 1996 ; 276 (8) : 637-9.

Les comités indépendants de surveillance des essais thérapeutiques : rôle et modalités de fonctionnement

B. Asselain, A. Kramar

Un comité indépendant de surveillance d'un essai thérapeutique (IDMC pour *Independant Data Monitoring Committee*) est constitué d'un groupe d'experts indépendants, extérieurs à l'essai, soumis à une stricte confidentialité, dont le rôle est de suivre la progression de l'essai, de s'assurer de la sécurité *(safety)* des patients qui y participent et, lorsque cela est prévu dans le protocole, de décider, à partir des résultats des analyses intermédiaires d'efficacité, de la poursuite ou de l'arrêt de l'essai en fonction de règles préétablies.

Mis en place avant tout pour des raisons éthiques, ces comités ont également pour mission de veiller à la bonne qualité du monitoring, d'examiner d'éventuels amendements proposés par le comité de pilotage ou le promoteur et de proposer d'éventuelles modifications du protocole.

Ils sont particulièrement nécessaires au cours des grands essais de phase III en cancérologie, essais qui nécessitent souvent une longue durée d'observation des patients, mais ils peuvent être également d'une grande utilité dans les essais de phase II, en particulier les essais de phase II randomisés.

De fait, les comités de surveillance ont la possibilité de lever l'aveugle et de formuler des recommandations qui auront un impact sur la conduite de l'essai. L'accès à de telles données peut entraîner l'interruption de l'essai avant son terme et la responsabilité d'une telle décision doit être longuement réfléchie.

En effet, devant des résultats précoces souvent impressionnants il est vrai, plusieurs grands essais ont été interrompus à partir d'analyses très préliminaires, conduisant souvent à changer le traitement des patients du bras de référence pour le produit actif, rendant impossible par la suite toute analyse à long terme de l'essai [1].

Nous allons donc tenter de répondre à une série de questions pratiques que pose la mise en place d'un IDMC.

Quand mettre en place un comité de surveillance ?

Durant la planification de l'essai, le promoteur et le comité de pilotage doivent évaluer l'intérêt de la mise en place d'un IDMC.

Seront pris en compte le type de l'essai, le critère de jugement, la durée de l'étude et, bien sûr, la connaissance que l'on a *a priori* sur la tolérance et la toxicité des médicaments à l'étude.

Dans le cadre des maladies comme le cancer, qui mettent en jeu le pronostic vital et où les traitements efficaces sont en général toxiques, la mise en place d'un tel comité est recommandée.

Si des analyses intermédiaires permettant un arrêt précoce de l'essai – que ce soit pour efficacité ou pour inefficacité – sont prévues dans le protocole, la mise en place d'un comité indépendant est pratiquement obligatoire pour assurer la rigueur et la crédibilité du processus de décision. Ce comité intervient non seulement pour donner un avis en faveur ou en défaveur du traitement expérimental, mais il est de plus en plus sollicité pour des décisions concernant la futilité *(futility)*, pour pouvoir arrêter un essai qui ne sera pas concluant.

Dans certaines situations, en particulier si l'essai est conduit dans un laps de temps très court, sans analyse intermédiaire, et lorsque la toxicité des drogues testées est parfaitement connue, alors il n'est pas nécessaire de mettre en place un comité indépendant, qui alourdit la gestion de l'essai. Cela peut être le cas de certains essais de phase II, où des comités indépendants d'évaluation de la réponse peuvent être beaucoup plus utiles qu'un comité de surveillance de l'essai, même si on ne peut mettre ces deux comités sur un même plan.

Quelles sont les responsabilités d'un comité de surveillance ?

Le comité de surveillance émet un avis consultatif, les décisions étant prises par le comité de pilotage de l'essai et par le promoteur.

Sa première tâche est de s'assurer de la qualité de l'essai : respect du rythme des inclusions, observance du traitement, déviations par rapport au protocole sont des indicateurs essentiels à surveiller afin d'alerter le comité de pilotage en cas de problème dans la conduite de l'essai [2, 3].

Dans la majorité des cas, le suivi de la sécurité des patients constituera la mission essentielle du comité de surveillance. Il surveillera donc les événements indésirables, dans leur ensemble, et analysera plus particulièrement les événements indésirables graves (SAE pour *Serious Adverse Events*). Un envoi mensuel des événements indésirables graves à l'ensemble des membres du comité est souvent proposé.

Si le comité peut travailler « en aveugle » dans un essai randomisé, il peut à tout moment – s'il l'estime nécessaire – lever cet aveugle pour avoir une meilleure appréciation du rapport bénéfice/risque de l'essai afin de pouvoir pondérer un éventuel excès de risque par un avantage potentiel.

Il peut également demander, en cas de toxicité inattendue, l'accès aux données sources du patient et requérir l'avis de spécialistes non représentés au sein du comité. Il recherchera en particulier d'éventuelles interactions médicamenteuses. En cas de toxicités jugées trop sévères, des recommandations peuvent être proposées : modification des critères d'inclusion, modification des traitements, arrêt temporaire des inclusions, voire arrêt définitif de l'essai ou de l'un des traitements de l'essai.

En cas de rythme d'inclusion trop lent par rapport aux prévisions, l'IDMC peut émettre un avis sur les mesures correctives entreprises par le comité de pilotage.

L'IDMC jugera de la qualité des données qui lui sont présentées, car il ne peut délibérer valablement que sur des données actualisées reflétant au mieux l'état des patients au moment de la réunion du comité.

Un retard dans le monitorage rend évidemment illusoire une prise de décision adaptée et réactive du comité de surveillance. Cela est particulièrement vrai si l'IDMC a également pour mission d'appliquer des règles d'arrêt lors des analyses intermédiaires préspécifiées dans le protocole, règles qui doivent bien sûr avoir été acceptées et discutées avant le début de l'essai.

Dans de nombreux essais en cancérologie, il faut souvent attendre plusieurs mois pour qu'un monitorage efficace soit mis en place et il n'est pas rare que, lorsqu'arrive la date de la première analyse intermédiaire, le retard dans le recueil et la saisie des données soit tel que l'analyse ne peut être conduite dans de bonnes conditions. Il est alors de la responsabilité du comité d'exiger du promoteur un monitorage de l'essai de qualité, lui permettant d'effectuer réellement sa mission.

Le comité de surveillance doit également émettre un avis sur les amendements proposés par le comité de pilotage et le promoteur de l'essai.

Il doit également suivre l'évolution des essais similaires, car l'arrêt précoce d'un autre essai voisin peut avoir un impact scientifique et économique non négligeable sur la conduite de l'essai en cours, comme cela a été le cas au cours des essais de phase III des inhibiteurs de l'aromatase dans le traitement adjuvant du cancer du sein.

Quelle doit être la constitution du comité de surveillance ?

Les membres de l'IDMC sont choisis en fonction de leur expertise dans le champ de l'investigation concernée et/ou des toxicités spécifiquement attendues. Ils doivent avoir une expérience des essais cliniques et ne doivent pas avoir de conflit d'intérêt avec le promoteur ou les investigateurs de l'essai. La présence d'un statisticien est indispensable, surtout en cas d'analyses

intermédiaires d'efficacité ou de futilité. La présence d'un clinicien spécialiste du domaine de l'essai est également nécessaire. Un pharmacologue complétera ce « noyau » central du comité. En fonction des toxicités d'organe attendues, des spécialistes d'organes pourront compléter la composition du comité : cardiologue, neurologue...

Au total, l'IDMC sera composé au minimum de 3 personnes et ne dépassera pas 7 membres, afin de ne pas rendre les aspects opérationnels trop compliqués. Un nombre impair de membres permettra de prendre une décision à la majorité simple sans ambiguïté. Si le nombre de membres du comité est pair, la voix de son président sera prépondérante en cas d'égalité des votes.

Il faut en effet que le comité puisse parfois se réunir rapidement, éventuellement sous forme d'une conférence téléphonique, afin d'échanger sur un événement indésirable grave ou un projet d'amendement.

Comment fonctionne un comité de surveillance ?

Les règles de fonctionnement de l'IDMC devront être formalisées parallèlement à la finalisation du protocole pour être en cohérence avec ce dernier. Il est souhaitable de formaliser ce fonctionnement sous forme d'un document ou « charte de fonctionnement », qui sera établie par le promoteur et le comité de pilotage, et soumise à l'approbation de l'IDMC [4].

Le comité doit être pleinement fonctionnel avant l'inclusion du premier patient dans l'étude, et il est souhaitable qu'une première réunion commune avec le comité de pilotage et le promoteur soit organisée précocement afin de valider les règles de fonctionnement et finaliser la charte de l'IDMC. Cette charte devra couvrir les aspects administratifs, opérationnels, et méthodologiques.

Elle contiendra les sections suivantes :

- description de l'essai et de la méthodologie retenue, en particulier la méthodologie des analyses intermédiaires ;
- description des responsabilités de l'IDMC dans le cadre de cet essai ;
- liste des membres de l'IDMC avec leur qualification et leurs coordonnées ;
- fréquence et organisation des réunions « fermées » de l'IDMC ;
- règles de fonctionnement : *quorum*, élection d'un président parmi les membres du comité, rappel des règles de confidentialité ;
- description des procédures de communication avec le promoteur, le comité de pilotage, le centre de traitement des données statistiques : en pratique, le comité de surveillance émettra un avis à destination du comité de pilotage, qui devra être simple et clair, mais il gardera confidentiel le compte rendu de la réunion qui reflétera les débats ;
- format des documents dont le comité aura besoin lors de ses réunions : tableaux de toxicités, analyses d'efficacité, procédures statistiques, possibilité de demander des analyses supplémentaires ;
- possibilités de réunions « ouvertes » avec le comité de pilotage et le promoteur.

Comment se déroule en pratique une réunion du comité ?

La réunion commencera souvent par une partie « ouverte » au cours de laquelle le promoteur et l'investigateur principal informeront l'IDMC du déroulement de l'essai, du rythme des inclusions, de l'avancement du monitorage et des questions qu'ils se posent ou souhaitent poser au comité.

La deuxième partie, « fermée », aura lieu avec les seuls membres du comité et le statisticien chargé de l'analyse des données. Ce statisticien peut être le statisticien du comité, mais ce n'est pas souhaitable et il est préférable que les données présentées le soient par un statisticien indépendant à la fois du promoteur et de l'IDMC. Il sera évidemment soumis aux mêmes règles de confidentialité que les membres de l'IDMC.

Ce statisticien présentera les résultats de l'essai : tableaux de toxicité, analyse des événements indésirables, analyse d'efficacité si celle-ci est prévue et rappel des règles de décision statistiques si des règles formelles d'arrêt ont été prévues.

Dans une troisième étape, les membres du comité discuteront entre eux des données présentées, rédigeront ensemble un avis qui sera transmis au comité de pilotage et au promoteur. Cet avis devra être le plus factuel possible avec des arguments pertinents. Il ne doit filtrer que des informations utiles pour la suite de l'essai.

Si tout se passe bien, que la toxicité est acceptable et qu'il n'y a pas d'élément indiquant que le traitement est délétère, l'avis sera souvent laconique : « Au vu des données examinées ce jour, l'essai peut continuer sans modification selon le protocole initialement prévu ».

Ces recommandations sont particulièrement importantes si le comité doit prendre des décisions formelles liées aux analyses intermédiaires, où la décision doit être prise bien sûr au vu des règles statistiques préétablies, mais également en prenant en compte les conséquences scientifiques et éthiques de ces décisions.

L'un des membres du comité, le président ou l'un des membres désigné secrétaire de séance, rédigera un compte rendu de la réunion, confidentiel, qui sera validé par les seuls membres de l'IDMC et gardé pour la fin de l'essai. Ces comptes rendus seront exigés au même titre que les avis rendus par les autorités de santé en cas de soumission du dossier pour enregistrement.

Quelles seront les implications des avis du comité sur l'analyse de l'essai ?

La décision d'arrêter un essai avant son terme doit être soigneusement pesée, car elle a bien sûr des implications importantes pour les patients de l'essai, mais aussi pour les autres essais qui posent des questions similaires.

Si l'IDMC a en charge le monitorage des analyses intermédiaires d'efficacité portant sur le critère de jugement principal de l'essai, la gestion du risque de faux positif (risque alpha ou erreur de type I) se pose de façon évidente [5]. Les méthodes de gestion du risque alpha sont maintenant bien connues (méthodes séquentielles groupées), les plus utilisées étant la méthode de Peto [6], qui propose de réaliser les analyses intermédiaires au seuil de 0,001 afin de conserver pratiquement intact le seuil de 0,05 pour l'analyse finale, et la méthode de O'Brien et Fleming [7], où le risque est dépensé progressivement au cours de l'essai, afin de garder également un seuil proche de 5 % pour l'analyse finale. La fonction de dépense du risque alpha proposée par Lan et Demets a l'avantage de la flexibilité [8], en dépensant le risque alpha à la demande, en fonction de la quantité d'information accumulée dans l'essai. La méthode retenue doit être précisée dans le protocole et validée avant le début de l'essai par l'IDMC.

Dans certaines circonstances, il se peut que l'IDMC ait à prendre en compte les résultats d'une analyse non planifiée du critère principal de jugement, par exemple pour évaluer si un surcroît de toxicité peut être compensé par un gain en termes d'efficacité. Il faut alors être conscient que ce type d'analyse a une influence sur le risque d'erreur de type I et prendre en compte l'impact de cette analyse sur l'analyse finale.

Dans certains plans expérimentaux adaptatifs, l'IDMC peut avoir à se prononcer sur une modification du nombre de patients à inclure en cours d'essai, et il faut alors que les règles du jeu soient clairement décrites dans le protocole. Là encore, des procédures appropriées de maîtrise du risque d'erreur de type I doivent être utilisées.

Quand aux analyses de futilité, permettant un arrêt précoce parce que les objectifs initialement fixés ne pourront vraisemblablement pas être atteints, elles impactent principalement l'erreur de type II et ont donc une influence sur la puissance de l'essai.

Conclusions

La généralisation des comités de surveillance dans les essais est assurément un progrès pour la protection des patients et pour l'éthique de la recherche clinique. Mais s'il est souvent important de pouvoir arrêter un essai précocement, il est également indispensable que l'essai puisse apporter une réponse claire et non ambiguë à la question qui était posée.

De nombreux essais ont été arrêtés prématurément ces dernières années en cancérologie, comme le critiquaient J. Bryant et N. Wolmark dans un éditorial du *New England Journal of Medicine* [9], en se basant sur des analyses certes contrôlées sur le plan statistique, mais avec un suivi insuffisant. On oublie souvent que les règles d'arrêt sont calculées sous l'hypothèse d'une proportionnalité des risques au cours du temps, pas toujours vérifiée dans la pratique, et l'on a souvent vu « fondre » les risques relatifs associés à l'effet du traitement lors des mises à jour de l'essai.

L'indépendance de l'IDMC doit être réelle, son rôle est de veiller à la protection des patients et au bon fonctionnement de l'essai pour permettre d'aboutir à une réponse à la question principale

de l'essai. Il faut ainsi savoir résister aux tentatives de demande d'information que ne manqueront pas de faire les investigateurs ou le promoteur, et la première qualité d'un membre d'IDMC sera de rester muet « comme une tombe » face à ces interrogations pourtant apparemment légitimes.

La confidentialité stricte des informations détenues est une règle sur laquelle ne jamais transiger. Le comité veillera toujours à ce que ses avis soient suffisamment clairs pour ne pas être interprétés dans un sens ou dans un autre. Plus l'avis rendu sera court, moins il donnera lieu à des interprétations « sauvages » de la part des destinataires de l'avis.

Références

1. Goss PE, Ingle JN, Martino S, *et al.* A randomized trial of letrozole in postmenopausal women after five years of tamoxifen therapy for early-stage breast cancer. *N Engl J Med* 2003 ; 349 : 1793-802.

2. Brun-Buisson C. Les DSMB : quel rôle, quelles responsabilités ? *Médecine/Sciences* 2005 ; 21 : 78-82.

3. European Medicines Agency. Committee for medicinal products for human use (CHMP): Guidelines on data monitoring committees. *Statist Med* 2006 ; 25 : 1639-45.

4. DAMOCLES Study Group. A proposed charter for clinical trials data monitoring committees: Helping them to do their job well. *Lancet* 2005 ; 365 : 711-22.

5. Kramar A, Paoletti X. Analyses intermédiaires. *Bull Cancer* 2007 ; 94 (11) : 965-74.

6. Peto R, Pike MC, Armitage P, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient I. Introduction and design. *Br J Cancer* 1976 ; 34 : 585.

7. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979 ; 35 : 549-56.

8. DeMets DL, Lan KK. Interim analysis: The alpha spending function approach. *Stat Med* 1994 ; 13 : 1341-52 ; (discussion 1353-6).

9. Bryant J, Wolmark N. Letrozole after tamoxifen for breast cancer. What is the price of success? *N Engl J Med* 2003 ; 349 : 1855-7.

Plan statistique, rapport d'analyse statistique et rapport final d'essai

S. Gourgou-Bourgade, E. Mouret-Fourme, A. Savignoni

Le plan d'analyse statistique

Le plan d'analyse statistique (PAS) est un document de liaison essentiel entre la conduite de l'essai clinique et le rapport d'étude clinique. Les organismes réglementaires du Canada, de l'Europe, du Japon et des États-Unis s'attendent à ce que le PAS réponde aux exigences dans la spécification d'analyses déductives et autres techniques statistiques importantes. De plus, les laboratoires pharmaceutiques s'attendent à ce que le PAS fournisse des recommandations explicites à suivre par le statisticien. Le PAS est une étape distincte dans une étude de recherche clinique et il doit être approuvé avant le démarrage de la recherche. Il définit l'ensemble des analyses statistiques à exécuter dans le cadre de l'essai clinique et toutes les sorties logicielles (SAS ou autres logiciels) qui pourront être incluses dans le rapport d'essai clinique final (CTR pour *Clinical Trial Report*).

Le PAS est une section clairement identifiée dans le protocole et doit être approuvé avant le démarrage de l'étude clinique. Il doit contenir le détail des analyses statistiques planifiées dans le cadre de l'essai. Les analyses planifiées sont issues du résultat des productions attendues et doivent être conduites de façon cohérente et répétable.

Le PAS doit contenir les conditions et les paramètres détaillés des résultats rapportés de l'essai, le format et le contenu des rapports produits et les tests utilisés pour soutenir la robustesse et la sensibilité de l'analyse conduite.

Les contrôles recommandés et des procédures spécifiques dans la mise en œuvre et l'achèvement de l'analyse permettant d'assurer la qualité des données et empêchant la compromission de l'étude doivent être inclus.

Le PAS doit inclure, au minimum, pour chacun des critères (principal et secondaires) :
- la définition précise du critère ;
- comment le résultat sera mesuré ;
- les transformations requises sur les données d'origine avant analyse ;
- les tests statistiques appropriés utilisés pour l'analyse des données ;
- comment les données manquantes seront gérées et présentées dans les analyses ;

- le cas échéant, les méthodes d'ajustement statistique réalisées.

Le PAS décrit point par point ce que le rapport statistique va contenir et qui est décrit plus bas.

Le rapport d'analyse statistique

L'objectif de ce chapitre est de présenter les recommandations quant au contenu d'un rapport statistique d'essai clinique, ainsi que les procédures pour sa préparation, sa révision et son approbation. Ce chapitre a été construit essentiellement à partir des textes de référence connus [1-3] et des expériences des auteurs. Il s'agit de suggestions pour un contenu minimal.

Définitions

Compte tenu de l'importante production scientifique issue des résultats des essais cliniques, des recommandations internationales [1] ont vu le jour dans les années 1990. Il s'agissait ainsi de promouvoir une standardisation du report des données d'essais cliniques en termes de qualité et de fiabilité de ces résultats. Ces recommandations mettent en avant l'importance du contenu minimum du rapport d'essai clinique et plus particulièrement sa partie statistique.

Le rapport statistique est essentiel car il constitue le document de référence pour la traçabilité de l'ensemble des données et de leur analyse, gage de qualité en vue notamment de l'exploitation statistique. Il comprend une description détaillée des résultats de l'essai clinique et des interprétations statistiques de ces résultats.

Plus, globalement, le rapport statistique fait partie intégrante du rapport final d'essai clinique, qui comporte également l'interprétation clinique des résultats ainsi que les aspects réglementaires relatifs à la vie de l'essai clinique.

Les acteurs

Le rapport statistique doit être rédigé par le statisticien responsable de l'essai ou sous sa direction en collaboration avec l'investigateur coordonnateur de l'essai. L'ensemble du document doit être validé et approuvé par le promoteur ou son représentant avant toute diffusion des résultats. Un formulaire recueillant les signatures est associé et archivé avec le rapport. Toute modification devra faire l'objet d'un circuit de relecture et de validation identique.

Les différents types de rapport

Le rapport statistique est rédigé en accord avec le PAS. Il peut être rédigé lors d'une analyse intermédiaire (tolérance ou efficacité) pour la réunion d'un comité de surveillance et pour l'analyse finale de l'étude. Le contenu dépend du type de rapport et il doit être adapté à son objectif. Il doit être systématiquement associé à un gel de la base de données en collaboration avec le gestionnaire de données *(data-manager)*.

Le rapport statistique sera le support ou pourra être inclus dans le rapport d'essai clinique final sous la trame proposée par les recommandations internationales de l'*International Conference on Harmonisation – Efficacy 3* [2].

Le contenu minimal

Le rapport statistique devrait contenir systématiquement les paragraphes suivants dans le corps du texte ou ses annexes et sera adapté selon l'étape à laquelle il sera élaboré.

Résumé des résultats

Cette section présente le résumé de l'ensemble de la recherche et de ses principaux résultats (titre, objectifs, population étudiée, description du traitement, méthodologie, hypothèse et analyse statistique, durée de l'essai, résultats principaux des analyses intermédiaires le cas échéant, l'observance, l'efficacité, la tolérance, la qualité de vie, conclusions).

Liste des abréviations et définitions des termes utilisés

Cette section présente la liste des abréviations et définitions des termes utilisées dans le rapport.

Éthique et réglementaire

Cette section présente le récapitulatif des différentes autorisations des autorités compétentes [comité de protection des personnes (CPP) et Agence française de sécurité sanitaire des produits de santé (Afssaps) et éventuels amendements (joindre en annexe les copies)]. Cette partie est obligatoire dans le rapport final d'essai clinique.

Investigateurs et structure administrative de l'étude

Cette section liste les investigateurs et leur lieu de travail et coordonnées, identification de la cellule de gestion de l'essai clinique. Cette partie est obligatoire dans le rapport final d'essai clinique.

Introduction du rapport

Il s'agit de présenter un rappel du rationnel accompagné d'une description du plan de l'essai, des objectifs principaux et secondaires du protocole, de la méthodologie pour le calcul du nombre de sujets nécessaires et des méthodes de randomisation éventuelles de façon succincte.

Critères d'inclusion et de non-inclusion

Cette section présente la liste des critères d'inclusion et de non-inclusion.

Critères d'efficacité

Cette section présente une définition détaillée du critère de jugement et la méthode d'évaluation. Par exemple, dans les essais de phase II, on utilise les critères RECIST [4] ou des critères de survies dans les essais de phase III.

Critères de tolérance

Cette section présente une définition détaillée des critères étudiés pour évaluer la tolérance. L'échelle de cotation utilisée et sa version seront nécessaires : par exemple, CTC-NCI version 4 [5], des indicateurs en gériatrie, l'évaluation de la qualité de vie (QLQ-C30) [6].

Assurance qualité

Cette section présente une définition du scénario de saisie réalisé (simple ou double), les logiciels utilisés pour la gestion des données et pour l'analyse statistique. Il s'agit de référencer les documents précisant les contrôles de qualité effectués sur la base de données et les différentes modifications intervenues au cours de l'étude.

Traitement des données et méthodologie statistique

Cette section présente une description et justification de toutes les variables générées et transformées dans l'analyse ; définition des groupes d'analyse [ITT *(Intention To Treat)*, ITT modifié, per-protocole, tolérance] ; traitement des données manquantes (fréquence et méthode de prise en charge) ou des données problématiques (patients perdus de vue) ; méthode de levée d'aveugle, définitions des déviations majeures, mineures et écarts au protocole et méthode de traitement des différentes déviations ; détails de toutes les déviations par rapport au PAS ; détails et justification des tests statistiques, des hypothèses nulle et alternative associées au critère principal, des niveaux de signification et des techniques d'estimation, effets des analyses intermédiaires, méthodes pour la prise en compte des comparaisons multiples. Le recours à l'utilisation d'analyses statistiques complexes, telles que les analyses multivariées, devra être décrit de manière précise.

Population étudiée

Cette section présente une description détaillée de la population étudiée. La courbe des inclusions théoriques et observées sera présentée, la période d'inclusion sera décrite ainsi que le nombre d'inclusions par centre et le rythme moyen *(figure 1)*.

Le nombre de patients recrutés, randomisés, traités, évaluables pour la tolérance et évaluables pour l'efficacité sera détaillé pour chaque bras et sera résumé sous forme de tableau. L'exemple présenté dans le *tableau I* concerne 150 patients (75 par bras) qui ont été randomisés et constituent donc la population en intention de traiter. Un patient n'a pas été traité dans le bras A et il n'est donc pas évaluable pour la toxicité. Un autre patient dans le bras A n'a pas été évalué pour l'efficacité.

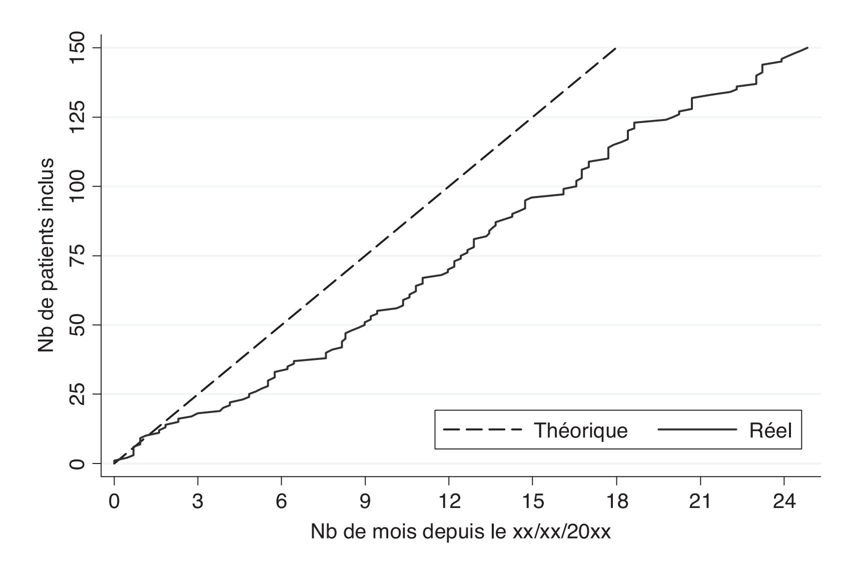


Figure 1. Exemple de courbe d'inclusion.

Tableau I. Exemple de tableau présentant les patients pour les deux bras de traitement.

Patients	Bras A		Bras B		Total	
	N	%	N	%	N	%
Inclus/randomisés	75	100	75	100	150	100
Population ITT	75	100	75	100	150	100
Traités	74	98,7	75	100	149	99,3
Évaluables pour la tolérance	74	98,7	75	100	149	99,3
Évaluables pour l'efficacité	73	97,3	75	100	148	98,7

ITT : intention de traiter.

Les raisons des retraits de consentement devraient être précisées ainsi que toutes les déviations au protocole et la justification des différences de population évaluables pour la tolérance et pour l'efficacité.

Un diagramme récapitulatif des inclusions selon l'énoncé CONSORT [1] sera associé à la description de la population étudiée. Pour établir ce diagramme (*flowchart* ou *trial profile* dans certaines publications), les informations suivantes sont nécessaires : le nombre de patients éligibles, non sélectionnés (et les raisons), randomisés, ayant reçu le traitement alloué par la randomisation (et les raisons sinon), perdus de vue/non traités/ayant le traitement interrompu, analysés (et les raisons sinon). Ce diagramme sera nécessaire à la publication *(figure 2)*.

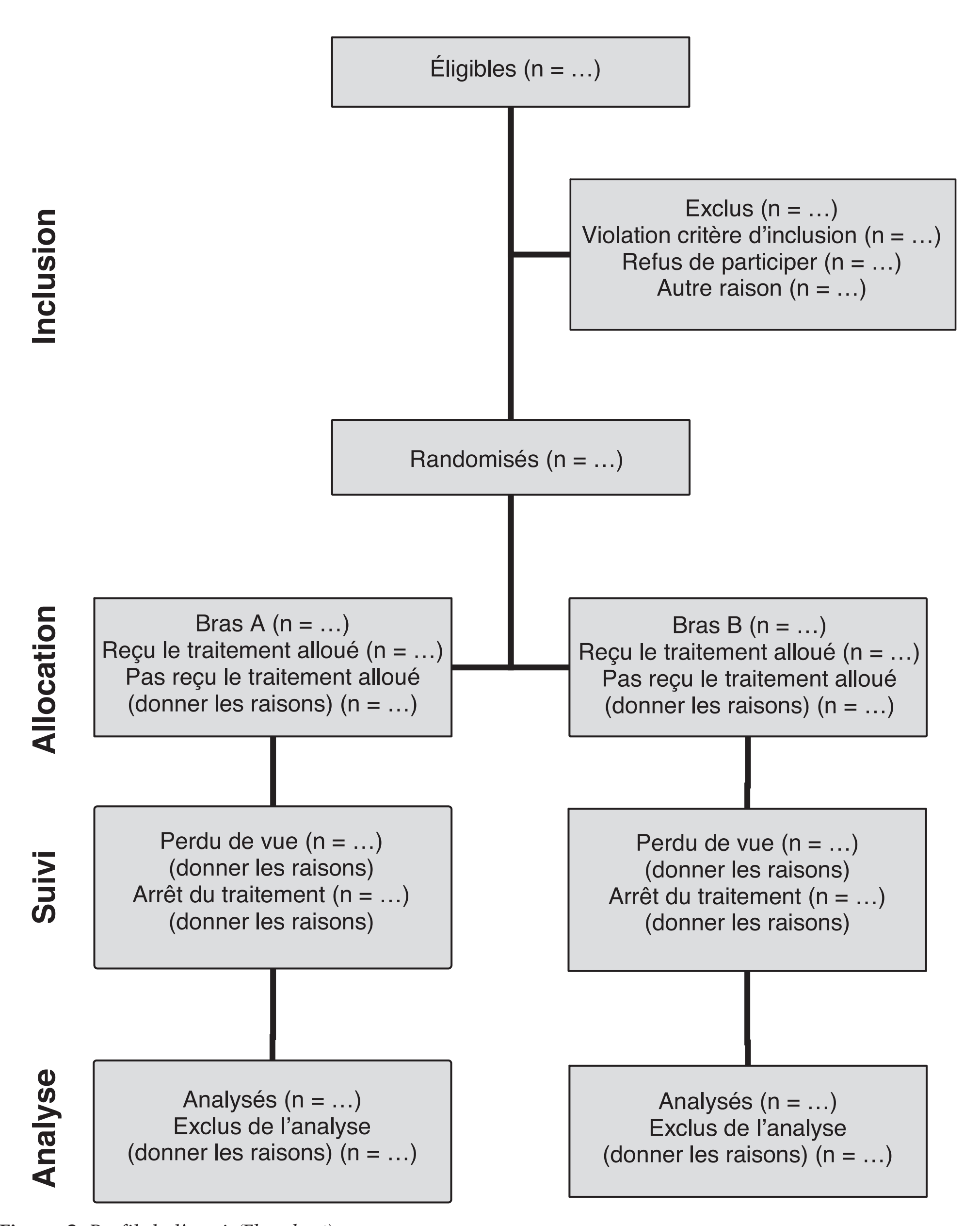


Figure 2. Profil de l'essai *(Flowchart)*.

Résultats

Cette section présente les caractéristiques démographiques (âge, sexe, etc.), cliniques (taille de la tumeur, envahissement ganglionnaire, stade, etc.) et biologiques initiales. La comparabilité des caractéristiques initiales entre les groupes sera présentée par des moyennes ou pourcentages selon la nature de la variable (*cf.* chapitre I.1 « Distributions statistiques », page 3).

Concernant les essais comportant de la chimiothérapie, par exemple, la tolérance sera décrite à la fois par patient (grade maximal) et par cycle (grade). L'observance du/des traitements sera évaluée par type de traitement en précisant le nombre total de cycles, les délais moyens inter-cycles, les doses (dose prévue, dose totale administrée, dose intensité et dose-intensité relative), les retards de plus de 7 jours par exemple (selon le schéma d'administration), ainsi que les réductions de dose et leurs causes. Les arrêts prématurés seront dénombrés et leurs causes explicitées.

Les événements indésirables graves (EIG) seront dénombrés et listés (en termes d'EIG « attendus » et « inattendus ») et doivent être en accord avec les déclarations aux autorités compétentes suite à la réconciliation avec la base de données de pharmacovigilance.

L'efficacité sera décrite en fonction de la méthodologie préalablement définie et présentée de façon adéquate. L'analyse de l'efficacité pourra être réalisée par groupe de patients « randomisés (ITT) » et/ou « évaluables pour l'efficacité (per-protocole) » et/ou « traitement reçu » (notamment pour la tolérance). En cas d'analyse intermédiaire, le rapport final doit présenter l'ensemble des résultats (un résumé de l'analyse intermédiaire et une analyse finale sur le critère principal).

L'ordre de présentation des résultats sera déterminé par les objectifs du protocole :
- *pour un essai de phase II :* les taux de succès (réponse tumorale, survie sans progression, par ex.) seront présentés avec leur intervalle de confiance à 95 % ou 90 % selon le risque alpha choisi dans le plan expérimental pour le critère principal de l'essai ;
- *pour un essai de phase III :* les résultats suivants seront présentés : le suivi médian (*cf.* chapitre III.4 « Suivi et surveillance », page 181) avec sa méthode de calcul, les probabilités de survie (globale, sans progression, sans maladie, par ex.) à des temps donnés accompagnés d'un intervalle de confiance (*cf.* chapitre III.1 « Données de survie », page 129) et le résultat du test statistique utilisé (*cf.* chapitre III.1), ainsi que le nombre et le type d'événements observés. Les graphiques présentant les taux de survie au cours du temps seront complétés par le nombre de patients à risque au cours du temps en bas de l'échelle des abscisses.

Dans le cas d'analyses multivariées, les résultats seront rapportés selon les recommandations détaillées dans le chapitre IV.2 (« Modèle de Cox et index pronostique », page 213).

Dans le cas d'analyses en sous-groupes, il sera nécessaire de préciser que ces analyses ont été effectuées à titre exploratoire et si elles avaient été prévues dans le protocole pour générer éventuellement des hypothèses à tester dans un futur essai (analyses en sous-groupe sur les covariables de stratification *a priori* et autres).

Discussion et conclusion

Cette section présente un résumé des principaux résultats statistiques avec leur interprétation. Des facteurs comme la qualité des données, des problèmes d'exécution de l'essai ou des problèmes extérieurs (temps) qui pourraient influencer les résultats devront être considérés. Les éventuels biais devront être discutés.

Cette partie doit être rédigée par l'investigateur-coordonnateur de l'essai et le statisticien doit s'assurer que les interprétations cliniques sont conformes aux résultats statistiques.

Tables, figures, annexes et liste de données

Tables

Cette section présente les tables de résultats qui doivent être identifiées, référencées dans le texte et indexées. Les unités de mesure, les explications des échelles d'évaluation et les groupes d'analyse sur lesquels les tables sont basées devront être indiqués dans le titre de la table.

Figures

Cette section présente les figures qui doivent être identifiées, référencés dans le texte et indexées.

Annexes

Cette section présente les différentes annexes qui doivent être identifiées, référencées dans le texte et indexées : le protocole, les avis des autorités compétentes, le cahier d'observation, le PAS.

Listes de données (Datalistings)

Cette section présente les listes de données individuelles qui doivent être identifiées, référencés dans le texte et indexées. Les *datalistings* sont la représentation des données enregistrées dans les cahiers d'observation (CRF pour *Case report Form*). Ils doivent être organisés par chapitre selon le CRF et être exhaustifs.

Données individuelles de patients (Patient profile)

Cette section présente certaines données individuelles de patients inclus, données majeures extraites et qui sont listées par patient sur une page unique (généralement pour les essais de phase I).

Conclusions

Dans la mesure du possible, le rapport statistique (ou les éléments statistiques du rapport clinico-statistique) doit être revu et approuvé avant la diffusion par un deuxième statisticien indépendant.

Le rapport statistique sera intégré ultérieurement dans le rapport final d'essai clinique comme recommandé par les guidelines internationales [2]. Le rapport statistique sera validé après signature d'un document traçant sa révision et son approbation. Ce document sera archivé avec l'ensemble des documents relatifs à l'étude.

Le contenu du rapport statistique est quasiment invariable mais peut-être adapté selon son objectif (rapport pour comité de surveillance, d'analyse statistique intermédiaire, rapport statistique final, rapport final d'essai clinique). L'utilisation d'un document standard peut guider et optimiser sa réalisation afin d'éviter l'omission de points importants. Sa validation pourra également être réalisée par la confrontation à la liste des chapitres à ne pas oublier selon l'énoncé CONSORT [1].

Ce rapport est essentiel car il permet d'avoir un historique complet de l'essai et constitue de ce fait le document de référence pour la rédaction de l'article scientifique final.

L'ensemble de ces recommandations nous permettent alors de répondre à l'obligation de fournir un rapport statistique qui est une demande explicite dans la loi française depuis 2006 suite à la transposition de la directive européenne pour l'application des bonnes pratiques cliniques [7].

Références

1. Moher D, Schulz KF, Altman DG ; for the CONSORT Group. The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001 ; 357 : 1191-4.

2. International Conference on Harmonization (ICH) E3: *Structure and content of clinical study reports*. December 95, CPMP/ICH/137/95.

3. International Conference on Harmonization (ICH) E9: *Statistical principles for clinical trials*. March 1998, CPMP/ICH/363/96.

4. Eisenhauer EA, Therasse P, Bogaerts J, *et al.* New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *EJC* 2009 ; 45 : 228-47.

5. Common Terminology Criteria for Adverse Events (CTCAE) and Common Toxicity Criteria (CTC). CTC-AE v4.0: http://ctep.cancer.gov/protocolDevelopment/electronic_applications/docs/ctcaev4.pdf

6. Aaronson NK, Ahmedzai S, Bergman *B*, *et al.* The european organization for research and treatment of cancer QLQ-C30: A quality-of-life instrument for use in international clinical trials in oncology. *J Natl Cancer Inst* 1993 ; 85 : 365-76.

7. Directive 2005/28/EC laying down principles and detailed guidelines for good clinical practice as regards investigational medicinal products of human use, as well as the requirements of authorization of the manufacturing or importation of such products. *Official Journal of the European Union* du 8 avril 2005.

Les logiciels

F. Kwiatkowski, E. Chamorey

L'avènement de l'informatique a complètement révolutionné l'usage des statistiques. Quand D.R. Cox a publié son article sur les modèles de régression en 1972, le calcul des résultats représentait un défi considérable tant en termes d'aptitudes que de temps. À cette époque, les plus gros calculateurs n'avaient pas la puissance de nos plus petits ordinateurs personnels. Aujourd'hui, utiliser un modèle de Cox ne représente plus une gageure, et nombre de chercheurs, grâce à la simplification de l'interface informatique, peuvent prétendre à l'utilisation de modèles sophistiqués sans même connaître les principes qui ont justifié leur développement et, malheureusement, sans non plus connaître les limites au-delà desquelles leur application devient incertaine (*cf.* chapitre IV.2 « Modèle de Cox et index pronostique », page 213). Dans cette évolution des techniques, nous aboutissons en fait à un paradoxe : parce que l'emploi des statistiques est désormais à la portée de tous, il s'est banalisé au point que beaucoup de chercheurs oublient le fait que des problèmes méthodologiques peuvent subsister en arrière-plan. La résolution des difficultés informatiques liées aux calculs a donné l'impression à l'utilisateur *lambda* que les questions méthodologiques ne se posaient plus, voire n'avaient jamais existé. Et aujourd'hui, c'est la méthode qui est la pierre d'achoppement. Les mises en garde dans ce domaine ne sont pas superflues car certains mythes relatifs à l'emploi des statistiques ont désormais cours ainsi que nous le rappelle Vickers [1] dans son article volontairement polémique : *How not to do research ?* Citons deux exemples : n'importe qui peut faire de la recherche clinique ; tout ce dont vous avez besoin pour faire des statistiques, c'est d'un bon logiciel (même si Excel fait déjà cela très bien).

Et l'auteur d'exhorter les biologistes et les médecins à ne pas réaliser des investigations sauvages et à faire appel aux statisticiens-méthodologistes, qui sont censés, eux, connaître les méthodes et les règles à suivre, et les pièges à éviter.

Après le rappel du contexte dans lequel nous évoluons aujourd'hui, ouvrons la mirifique boîte à outils statistiques qui s'offre à nous et jetons un coup d'œil (admiratif) sur l'immensité des ressources qu'elle met aujourd'hui à notre disposition. Évidemment, cette profusion de biens, nous la devons à la capitalisation des développements logiciels, permise non seulement par les progrès de l'informatique et de ses méthodes mais aussi et surtout, par la généralisation d'Internet qui a libéré les échanges et favorisé l'entraide (souvent bénévole) des développeurs et des chercheurs.

À retenir

À toutes les étapes de la recherche biomédicale, on trouve des outils parfois génériques, parfois spécialisés. En dépit de l'ergonomie de certains outils, l'intervention du statisticien demeure incontournable dans trois domaines principaux :
– dans la phase de mise au point du plan expérimental et ce, afin de garantir à l'expérimentation des chances de produire des conclusions avec un risque minimal de faux positifs ou de faux négatifs ;
– dans la phase finale d'exploitation des données permettant la vérification du bon déroulement du protocole (absence de biais), la production des résultats et leur interprétation ;
– désormais, comme la majorité des revues médicales à comité de lecture requièrent dans les auteurs la présence d'un statisticien, ce dernier intervient aussi dans la rédaction des publications avant la soumission.

Dans les grandes structures de recherche académique ainsi que dans l'industrie pharmaceutique, certaines tâches sont affectées à des personnels spécialisés, comme le data-management. Dans des structures plus modestes, une seule et même personne peut regrouper les fonctions de data-manager et de statisticien, parfois même d'assistant de recherche clinique (ARC), voire d'opérateur de saisie. Dans ces conditions, les logiciels utilisés peuvent différer et des applications effectuant les deux types de tâches (gestion des données et statistiques) peuvent être préférées. Chacun choisira les outils les mieux adaptés à ses activités spécifiques, sa régularité, son expérience et son environnement. D'autres tâches peuvent être externalisées pour des raisons de fiabilité et/ou de méthode, telle la randomisation.

Il va sans dire que l'expertise statistique s'étend parfois jusqu'à la proposition d'amélioration de tests statistiques existants, voire jusqu'au développement de nouveaux modèles, le plus souvent en parallèle avec la création de programmes dédiés.

Nous limiterons notre propos aux logiciels utilisés en statistique dans les différents secteurs d'activité des statisticiens œuvrant dans le domaine de la cancérologie. Nous les passerons en revue successivement en citant chaque fois que possible des logiciels libres disponibles, des routines de R (logiciel en accès libre) répondant aux questions ainsi que des sites Internet dans lesquels on peut trouver des programmes fonctionnels. Les logiciels commerciaux (SAS, SPSS, Stata, STATISTICA, etc.) ne seront pas cités, bien que l'on y retrouve toutes les routines souhaitées.

De nombreuses informations pourront être aussi trouvées chez Wikipedia, qui propose une documentation assez générale sur les statistiques aussi bien en français qu'en anglais (http://fr.wikipedia.org/wiki/Statistiques), mais cette encyclopédie montre des lacunes. Des sites, comme celui de l'Université Carnegie Mellon (Pittsburgh, États-Unis) http://lib.stat.cmu.edu ou celui de l'Université de Georgetown (Washington, États-Unis) http://statpages.org, référencent des centaines de logiciels libres et parfois des codes sources. Ils peuvent servir de base de recherche, mais ils peuvent faire perdre du temps par rapport aux logiciels commerciaux dans lesquels les fonctions sont organisées de manière structurée et où une aide interactive adéquate permet généralement d'aller rapidement au but.

D'autres sites comme celui de l'Université Vassar (Poughkeepsie, New York, États-Unis) http://faculty.vassar.edu/lowry/VassarStats.html sont plus conviviaux et offrent d'exécuter des fonctions en ligne directement sans passer par un téléchargement.

À retenir

On gardera à l'esprit toutefois que les citations de sites Internet ou de logiciels dans ce chapitre ne peuvent correspondre à une quelconque validation de notre part et, dans la mesure du possible, quand des chercheurs se servent d'utilitaires statistiques sur Internet, ils seront avisés de confirmer les résultats de leurs calculs par d'autres moyens. On se rappellera aussi que les sites ont une durée de vie limitée, qu'ils ne sont pas « testés » aussi robustement que les logiciels de référence et que leur actualisation ne suit pas toujours l'avancée des connaissances.

Définition du plan expérimental

Étape cruciale dans la recherche clinique mais aussi dans n'importe quelle expérimentation scientifique, elle a pour but, suite à la détermination des objectifs et des critères de jugement, d'optimiser les moyens employés, c'est-à-dire sur un plan statistique :

- de maximiser les chances d'aboutir à une conclusion (compte tenu d'hypothèses préalables) ;
- tout en minimisant les risques d'erreurs : celui d'énoncer une conclusion qui ne serait, en fait, pas conforme à la réalité (c'est-à-dire en concluant ou en ne concluant pas à tort) ;
- et ce, en réduisant au maximum l'exposition des patients à des traitements inférieurs de par leur efficacité ou leurs toxicités (aspect éthique).

Selon le type de plan expérimental, on trouve différents logiciels qui permettent, parfois grâce une succession d'itérations, de définir les conditions optimales pour répondre à la question énoncée dans l'objectif principal : la plus importante consiste en la taille de l'échantillon requis. Ce point est crucial car c'est de lui que dépendront les perspectives de succès du travail mais aussi les investissements à engager pour mettre sur pied l'expérimentation. Si l'effectif trouvé est trop important, sur un plan financier, les dépenses risquent d'être rédhibitoires ; sur un plan humain, la détermination d'un effectif minimal vise naturellement à réduire le nombre de patients soumis à un traitement n'ayant pas encore fait la preuve de son efficacité ou de son acceptabilité.

Essais cliniques

Le cas des essais cliniques est le mieux documenté. Les méthodes de calcul varient selon la phase des essais.

Essais de phase I

Une fois la détermination de la dose initiale et des paliers de doses faite (basée sur les études précliniques, voire cliniques si la molécule a déjà été testée sur l'homme), plusieurs méthodes

existent [2]. La méthode par paliers successifs de 3 individus ne nécessite pas d'outil logiciel particulier, la décision de passage d'un palier à l'autre se fait de manière empirique, selon le nombre de toxicités constatées. Cette méthode n'est toutefois plus recommandée par les méthodologistes, pourtant elle reste la plus utilisée en pratique [3]. Pour les autres méthodes, des calculs importants sont nécessaires et un appui informatique est indispensable dans le suivi et la prise de décision :

- la méthode avec réévaluation continue (CRM ou CRML) proposée par O'Quigley [4, 5] : cette méthode utilise l'approche bayésienne ou celle basée sur la maximisation de la vraisemblance. Le logiciel « np1 » traite ces deux approches [6] ;
- une autre approche bayésienne (EWOC) [7] est disponible en téléchargement gratuit ainsi que la documentation associée sur le site de l'Emory University à Atlanta : http ://sph.emory.edu/BRI-WCI.ewoc.html ;
- le *MD Anderson Cancer Center* de l'Université du Texas offre une grande variété de logiciels gratuits. On peut télécharger le logiciel « BMA CRM » qui utilise une approche CRM bayésienne pour la détermination de la dose maximale tolérée : http ://biostatistics.mdanderson.org/SoftwareDownload ;
- logiciel EPCT développé par David Machin au *National Cancer Center* de Singapour ;
- http://cran.r-project.org/web/packages/titecrm/index.html ;
- http://biostats.upci.pitt.edu/biostats/ClinicalStudyDesign/Phase1Standard.html ;
- http://www.cancerbiostats.onc.jhmi.edu/software.cfm.

Essais de phase II

Dans un essai de phase II, il faut d'abord choisir le bon plan expérimental *(design)*. Celui-ci induira, à partir des hypothèses d'efficacité ou de toxicité choisies, du nombre d'étapes, de l'éventuelle procédure de randomisation, un calcul du nombre de sujets nécessaires globalement et à chaque étape. Les règles de décision permettront de conclure et/ou d'interrompre l'essai soit pour conclusion positive, soit parce que le traitement s'avère insuffisamment efficace ou trop toxique (*cf.* chapitre V.2 « Mise en œuvre d'un essai clinique de phase II », page 281). Le site de l'Institut du Cancer de l'Université de Pittsburg permet d'effectuer les *designs* de Simon [8] et de Bryant & Day [9] : http://www.upci.upmc.edu/bf/resources.cfm.

D'autres méthodes multi-étapes [10, 11] peuvent également permettre une interruption plus précoce de l'essai en cas d'échec, les règles d'arrêt pouvant être déterminés à partir de nombreux logiciels (npII, de Rycke, SSS Project Machin 2009) [12].

D'autres approches bayésiennes ont été développées et programmées par les statisticiens du *MD Anderson* :

- pour des designs adaptatifs et un essai à un groupe *(Multc Lean)* ou à deux groupes *(Phase II PP)* évaluant à la fois l'efficacité et la toxicité ;
- pour la recherche de la dose optimale en regard de l'efficacité et de la toxicité *(EffTox)*.

Essais de phase III

La comparaison de deux traitements (ou plus) devrait toujours se baser sur des hypothèses assez solides, principalement celles issues des résultats de la phase II, voire ceux provenant d'autres

essais de phase III. On connaît l'efficacité du traitement de référence et on a une bonne estimation de l'effet du(des) nouveau(x) traitement(s) et des toxicités associées. La comparaison dépend de peu de paramètres : les risques de première et de deuxième espèce α et β (γ pour les essais de non-infériorité) et la différence attendue. Cette différence peut être exprimée à l'aide d'une variable qualitative (pourcentage de réponse), d'une variable quantitative (moyenne et écart-type) ou d'une variable censurée (taux de survie). L'implémentation des calculs dans Excel reste une solution pour qui connaît la formule adaptée, sinon de nombreux logiciels les proposent après vous avoir fait préciser le contexte de l'essai.

Le service de biostatistique du *Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins* (Baltimore, États-Unis) propose le téléchargement gratuit de plusieurs outils statistiques dont un nommé *Power* pour les calculs de puissance et d'effectifs des essais de phase III et ce, en fonction de différents types d'objectif : http ://cancerbiostats.onc.jhmi.edu/software.cmf (*remarque utile* : pour une utilisation en France de ce logiciel, il faut penser à configurer l'ordinateur avec le point décimal pour les nombres et non la virgule.)

Pour les essais intégrant les patients d'un essai de phase II dans une phase III, le *National Cancer Institute* (NCI) propose sur le site de la *Biometric Research Branch* une page permettant de calculer les effectifs de la phase III à partir des résultats de la phase II : http://brb.nci.nih.gov. Sur le même site, d'autres *designs* utilisant des biomarqueurs prédictifs de la réponse peuvent être modélisés.

Essais de phase IV

Les essais de phase IV ou de pharmacovigilance s'intéressent aux effets indésirables sur le moyen et long terme. En cancérologie, ils étaient jusqu'à récemment peu utilisés du fait de survies écourtées par la maladie. Néanmoins, pour certains cancers (sein, thyroïde, prostate, etc.), les survies sont désormais souvent supérieures à 15 ou 20 ans, et la pharmacovigilance et la pharmacoépidémiologie occupent une place plus importante. La découverte d'une augmentation de l'incidence de cancers de l'endomètre suite aux traitements hormonaux du cancer du sein [13] a été un des premiers éléments qui ont déclenché ce type de recherche. Les cancers radio-induits (leucémies) étaient eux aussi pointés du doigt depuis quelques années et ont été l'objet d'une attention particulière, surtout en pédiatrie. Nous n'avons pas trouvé de logiciel spécifique de cette phase qui puisse être conseillé.

Études rétrospectives

Tous comme les protocoles des essais cliniques, les protocoles des études rétrospectives doivent suivre une méthodologie précise et rigoureuse depuis le calcul du nombre de sujets nécessaires jusqu'à l'analyse statistique ajustée sur les facteurs potentiels de biais identifiés *a priori*. Dans ce type d'étude, la sélection des sujets peut poser des difficultés méthodologiques parfois incontournables et qui nécessitent une discussion avec un méthodologiste : par exemple, le nombre de patients, le tirage au sort à partir d'une base de données comme le choix d'une population témoin.

Très peu de logiciels proposent des routines permettant un appariement aléatoire sous contraintes. On se rappellera qu'en cas d'effectifs réduits, une des solutions consiste à utiliser un score de propension [14] et à apparier les patients sur ce score.

Inclusion des patients, affectation des traitements, gestion des données

Une fois le plan expérimental défini, les essais prospectifs de phase I à IV confrontent les investigateurs à un travail essentiel : l'inclusion des patients.

La plupart des logiciels de gestion d'essais commercialisés proposent des routines pour allouer de manière aléatoire des traitements. Ces fonctions prennent en charge les éventuelles stratifications. Comme nous l'avons dit en introduction, la randomisation peut aussi être externalisée, par exemple pour éviter que le tirage au sort soit entaché de partialité (par ex. un projet européen qui se termine en mai 2011 a permis le développement de ce type de fonction).

Enfin, le tirage au sort peut être volontairement pipé comme dans des stratégies de type « parier sur le gagnant » dans lesquelles on avantage le traitement qui semble donner les meilleurs résultats à mesure que les résultats s'accumulent dans l'essai. Certains logiciels commerciaux comme EAST (http://www.cytel.com) permettent la gestion de ce type d'approche.

Pour des essais incluant quelques centaines de patients ou plus, il serait aberrant de chercher des solutions en dehors des logiciels standards. Pour des recherches de plus petite taille ou dans l'expérimentation animale, des solutions alternatives peuvent être proposées. On pourrait faire appel à la fonction aléatoire ALEA() dans Excel. Par exemple, la formule ARRONDI.INF(ALEA()*3+1 ;0) renvoie un nombre aléatoire entre 1 et 3. Cela pourrait sembler idéal pour affecter des patients à 3 traitements différents. Hélas, cette fonction est « volatile », c'est-à-dire que son résultat se recalcule à chaque mise à jour de la feuille Excel et ce, quelles que soient les nouvelles données introduites. Le résultat doit donc être « figé », ce qui ne peut se faire que si on génère ces chiffres pour la série entière des patients à randomiser. Cela revient à réaliser une table de nombres au hasard, mais avec un défaut : une absence de blocs équilibrés. Dans ces conditions, mieux vaut utiliser une des tables fournies par les ouvrages de statistiques (par ex. Schwartz [15], en utilisant autant de blocs que de strates (*cf.* chapitre VI.2 « Modalités de randomisation », page 344).

Recueil des données ± gestion des essais

La gestion informatique des données de la recherche clinique représente à elle seule un ensemble considérable de solutions logicielles. Il va sans dire que les packages commercialisés sont assez coûteux et qu'ils répondent principalement aux besoins de l'industrie pharmaceutique. Les centres de lutte contre le cancer (CLCC) comme les grandes structures hospitalo-universitaires ou

hospitalières ne peuvent échapper à la standardisation des moyens et la professionnalisation des pratiques. Plusieurs éléments favorisent une telle évolution : l'optimisation de l'efficacité des moyens mis en œuvre, la recherche de la garantie de l'impartialité des recherches et, bien sûr, aussi la très grande taille des essais qui résulte de progrès attendus de plus en plus marginaux, la plupart du temps. Pour plus de détail, le lecteur peut se référer au chapitre VI.1 (« Gestion des données », page 331).

Analyses statistiques

Quel que soit le type ou la phase de l'essai, l'objectif est de produire des résultats à partir des données collectées. Parfois, ces résultats seront relativement pauvres, comme un simple dénombrement des réponses à des traitements, d'autres fois ils seront calculés avant même que l'essai soit terminé et ce, afin d'examiner l'utilité ou non de poursuivre l'expérimentation. En recherche clinique prospective, on fera généralement appel à des packages statistiques de référence tandis que pour des recherches plus ponctuelles, d'autres logiciels pourront suffire.

Dépendance entre système de gestion des essais et statistique

Étape incontournable avant l'analyse statistique, l'exportation des données. Aucun logiciel de gestion d'essais ne saurait exister sans un module de requêtes et d'export des données. De la même manière, un outil statistique ne sert à rien si on ne peut l'alimenter avec des données correctes. Généralement, divers formats d'export sont disponibles (.sas,. txt,. csv, etc.) et c'est la pratique qui guide les choix. Les tables exportées sont toujours rectangulaires, ce qui est une première limitation liée à la dichotomie logiciel de gestion/logiciel de statistique. Une autre provient de la structure de la base de données. Si elle n'a pas été convenablement bâtie, il sera éventuellement impossible d'extraire certaines informations à l'aide du module de requêtes, et donc de réaliser certaines statistiques ensuite. Enfin, les données généalogiques sont généralement intraitables quels que soient les logiciels employés.

Utilisation des packages de référence

Si pendant très longtemps le calcul statistique résumait à lui seul l'essentiel de la difficulté du travail des statisticiens, l'avènement de l'informatique a considérablement changé la donne. Effectuer des calculs ne pose plus guère de problème. Si l'essentiel du corpus statistique est finalisé, on peut légitimement penser que les logiciels actuels répondent déjà à la quasi-totalité des questions qui se posent en recherche clinique et qui se poseront encore demain. L'utilisation de ceux-ci, en particulier ceux de référence (SAS, SPSS, Statistica, Stata, StatGraphics, etc.), garantit la qualité des résultats dès lors que l'on a bénéficié de la formation adéquate. De la même manière, le logiciel libre R permet tout calcul statistique et ce logiciel soutient la comparaison avec les précédents.

Les logiciels commerciaux de gestion des essais (Capture-system, Clintrial, Macro) ne disposent que de peu de fonctions statistiques. Souvent les interrogations devront se limiter à de simples requêtes à des fins de dénombrement et c'est l'exportation des données vers un logiciel de statistique qui va permettre l'exploitation des résultats

Utilisation d'Excel

Pour des analyses rétrospectives simples (matrice rectangulaire de données tenant dans une feuille), on peut parfois se satisfaire des fonctionnalités statistiques d'Excel. Elles sont relativement développées dans la version de base et le didacticiel est de bonne qualité. Bien sûr, certaines tâches aussi banales que de vérifier les hypothèses de normalité, de tracer une courbe de survie... seront à proscrire. Les modèles multivariés ne sont, eux non plus, pas disponibles.

Des solutions à ces lacunes existent pourtant, grâce à l'adjonction de logiciels complémentaires payants comme XLStat, StatBox, StatTools. Ces logiciels ajoutent souvent une barre d'outils à la barre de menus d'Excel pour faciliter son fonctionnement et gérer les nouvelles options.

Enfin, pour des besoins particuliers et pour des utilisateurs avertis, il est possible de créer ses propres programmes en arrière-plan d'Excel (en Visual-Basic).

Analyses génomiques, transcryptomiques, protéomiques

Les analyses génomiques font appel à des calculs souvent plus mathématiques que statistiques. Mais vue l'importance que prennent les recherches dans ce domaine, il est nécessaire de se doter d'outils adaptés pour traiter les données, notamment pour le *clustering* (classification automatique hiérarchique). L'Institut du Cancer de l'Université de Pittsburgh [16] propose deux approches logicielles pour couvrir l'essentiel des besoins : une interface Internet pour effectuer le chargement des données, le calcul en ligne et le téléchargement d'un exécutable pour réaliser ces tâches sur son ordinateur. Ils sont disponibles à l'adresse : http ://bioinformatics2.pitt.edu/GE2/GEDA.html.

De la même manière, on peut utiliser le programme MeV [17] créé grâce à la *National Library of Medicine* américaine et le *Dana-Farber Cancer Institute* de Boston. C'est aussi un logiciel libre que l'on peut télécharger à partir de l'adresse : http://www.tm4.org/mev.

Ces deux logiciels sont tous deux très fréquemment référencés dans la littérature et peuvent être utilisés sans risque, dès lors qu'on a les compétences adéquates... On pourrait toutefois leur reprocher la non-implémentation la méthode de Ward dans les méthodes de *clustering* et l'absence de questionnement quant au risque α quand les analyses intègrent plusieurs milliers de gènes et seulement quelques dizaines de patients.

Publication des résultats

Cette étape pourrait en première analyse être ignorée dans ce chapitre « logiciels informatiques ». Toutefois, l'utilisation de logiciels particuliers par le statisticien est bien requise à cette étape, principalement pour présenter les résultats sous forme de graphes ou de figures qui seront ensuite « collés » dans les textes. Par ailleurs, les statisticiens sont aujourd'hui interpellés quant à des statistiques de bibliométrie sachant que les chercheurs sont évalués à l'aune de leurs publications… ainsi que le financement national de la recherche biomédicale dont une partie revient aux établissements de santé par le biais des enveloppes MERRI (mission enseignement, recherche, référence et innovation).

La production de figures

Depuis plus d'une décennie, il est devenu très aisé de produire un graphe, que ce soit pour des statistiques descriptives, des corrélations, des différences de moyennes ou encore des courbes de survie et des modèles multivariés. D'ailleurs, nombreux sont les logiciels qui produisent directement ces diagrammes en même temps que les résultats sous forme chiffrée.

Une première remarque est nécessaire à ce niveau, relative au format d'exportation de ces figures. Généralement, c'est un format bitmap (textuellement carte de points) très pratique pour le programmeur mais peu pour l'utilisateur. Ce format se reconnaît à l'extension du fichier (.BMP, .GIF, .JPG) qui implique que l'objet en question n'est pas facilement modifiable une fois exporté. Cela signifie que la résolution de la figure peut sembler suffisante quand on la regarde à l'écran, mais qu'elle peut être très mauvaise quand on la place sur un support imprimé (flou, grossièreté des contours, mauvaise qualité des chiffres et des lettres, etc.). Cela est particulièrement vrai pour les formats compressés comme le format JPEG qui perd en qualité ce qu'il gagne en volume. Si ce problème s'estompe du fait de la généralisation des revues *on-line*, il demeure pour la production de livres mais aussi pour les communications orales avec les diapositives souvent projetées en grand format devant les assemblées. Il y a deux manières de pallier ce problème :

- la première concerne le moment de la copie : si le logiciel permet de faire varier la taille de la figure à l'écran, il faut s'arranger pour enregistrer la représentation la plus large, car une image perd moins en netteté quand on la diminue que quand on l'agrandit. Si, par ailleurs, le logiciel permet un format d'export de très haute résolution, on gardera celui-là de préférence, sauf s'il s'agit d'envoyer la figure dans un e-mail par Internet ;
- la seconde consiste à éviter les formats compressés : le format BMP est gourmand mais il conserve l'entièreté de l'information. Et pour réduire la taille d'un BMP, on songera à réduire le nombre de couleurs – un format en 28 000 couleurs est une aberration – plutôt que de restreindre la taille de la figure.

La retouche de ces images n'est jamais facile. Microsoft fournit avec Windows un programme nommé PAINT – Apple, sur Mac, propose un logiciel similaire – qui permet de modifier n'importe quoi dans une figure au format BMP… mais le temps qu'il faut pour cela est proportionnel naturellement à l'étendue des modifications et à la connaissance du logiciel (par ex. comment mettre

à la verticale le titre des ordonnées ?). Dans ce type de logiciel, le plus chronophage toutefois est le recoloriage de lettres et de points non contigus. Il est possible d'agrandir ou de réduire une figure, mais cela se fait au détriment de sa qualité. En conclusion, si la retouche d'une figure ne demande que l'ajout/remplacement de certains libellés, c'est tout à fait possible. Sinon, ce n'est probablement pas une bonne solution.

Deux autres solutions permettent la production de figures de bonne qualité :

• Excel est capable de tracer de nombreux graphiques. On peut procéder en copiant seulement les tableaux de résultats issus des logiciels de statistique et utiliser le grapheur d'Excel qui dispose d'un choix souvent suffisant de types de graphe. Bien sûr, ceux qui savent programmer sous Excel (c'est-à-dire réaliser des routines en Visual-Basic) pourront produire une variété de graphes presque sans limite ;

• PowerPoint (ou tout autre logiciel de présentation) : si le graphe consiste en courbes, il est possible de le copier dans une diapositive, de l'agrandir à taille voulue en arrière-plan, puis de tracer par-dessus les nouvelles courbes avec l'option « formes automatiques, lignes et formes libres ». Une fois la figure achevée, on retire l'image de fond, on groupe les éléments et on peut leur donner la taille que l'on veut sans perdre en qualité – c'est un format vectoriel – et changer les couleurs comme bon nous semble.

Bibliométrie

Deux types de logiciels sont utilisés dans ce domaine :

• ceux permettant de récupérer les listes annuelles de publications des chercheurs sur Internet ;
• ceux produisant des statistiques à partir de ces listes.

Récupération des publications

Que l'on dispose ou non d'un service interne qui centralise les publications des chercheurs de l'établissement, il est indispensable d'utiliser les outils de recherche du web pour garantir l'exhaustivité du recueil. Entre autres, parfois certains auteurs ne savent même pas qu'ils figurent sur des publications originaires d'autres centres. Plusieurs solutions sont disponibles à cette fin :

• PUBMED : c'est le site américain http ://www.ncbi.nlm.nih.gov/pubmed/ bien connu des chercheurs. Y sont référencées les principales parutions dans le domaine de la biologie mais aussi certaines dans le domaine psychologique. Sont absentes nombre de revues franco-françaises et d'autres sans *impact factor*. Par ailleurs, si l'on veut être extrêmement réactif, on ne trouvera pas les articles publiés dans le mois précédent ;

• *ISI Web of Knowledge* (ou *Web of Science*) : http://apps.isiknowledge.com. Il est en particulier accessible aux personnels rattachés à l'Inserm. Un peu plus exhaustif que PUBMED, il permet des recherches groupées (pour x auteurs par ex.) assez efficaces. Sa limitation des articles affichés à 500 se révèle une limite assez rapidement atteinte quand on travaille au niveau d'un établissement de santé. Ses formats d'exportation facilitent les transferts vers d'autres bases de données. Enfin, il est le seul à donner le nombre de citations actualisé des articles, donnée indispensable si l'on veut calculer les indices G ou H des auteurs ;

• *Google Scholar* (http://scholar.google.fr/) : ce site effectue les recherches de manière très large, avec, entre autres, les publications de chapitres de livre, ce que ne font pas les deux sites précédents. Malheureusement, un pourcentage non négligeable des éléments retournés par une requête ne concerne pas les mots clés entrés : par exemple, en recherchant les citations d'un auteur, on pourra ramener des annonces publicitaires. Il possède cependant des avantages : il est très exhaustif et il est actualisé très rapidement. Il apparaît donc comme un bon complément aux travaux de bibliométrie quand ils portent sur un nombre restreint de personnes.

Publimétrie

En 2011, le premier référencement incontournable national français des articles médicaux est le SIGAPS (système d'interrogation, de gestion et d'analyse des publications scientifiques) : http://sigaps.fr.

Le score calculé par SIGAPS pour un établissement s'établit à partir de l'*impact factor* des revues où sont publiés les articles et de la position des auteurs. Malgré les habituelles limites de ce type de score – est-ce un vrai reflet de la qualité de la recherche ? –, selon une étude récente [18] le score SIGAPS est apparu « robuste » et peut servir de première base dans la comparaison des établissements de santé, des universités mais aussi des chercheurs à l'intérieur d'un établissement. On pourra cependant regretter que le nombre de citations n'intervienne pas dans ce mode de calcul, un article paraissant dans une revue de grande notoriété pouvant n'avoir que peu d'impact dans la réalité, ce qui transparaît par son nombre de citations. Il est possible d'envisager des indices de citation qui tiennent compte de l'*impact factor* des journaux où elles apparaissent.

L'*ISI Web of Knowledge* possède, outre son bon référencement des articles médicaux, quelques outils statistiques avantageux, en particulier le calcul aisé de l'indice H pour un auteur ou pour un établissement tout entier. Mais ce site ne gère pas les homonymies. Si un auteur s'appelle « Dupont », il va lui être très difficile d'utiliser sans difficulté les calculs de ce site, sinon à effectuer une requête croisée avec le nom de son établissement. Par ailleurs, ses statistiques sont relativement limitées, et le calcul de l'indice H, référence assez largement utilisée, n'en fait pas un exemple d'illustration de la production scientifique. Il faut donc principalement s'en servir comme source de données et effectuer ses propres statistiques autrement. À notre avis, le travail de centralisation des publications par chaque établissement de santé, rétrospectivement et prospectivement, est incontournable et n'est ni suffisant ni efficace si l'on se limite à SIGAPS ou à l'*ISI Web of Science*.

Conclusions

La boîte à outils statistiques idéale est celle qui permet de résoudre les problèmes de ses utilisateurs. Comme nous venons de le dire, la plupart des outils statistiques existent, assez facilement accessibles surtout si l'on utilise des logiciels de référence. Les utilisateurs de R qui est un logiciel

libre et dont nous conseillons l'utilisation, mais ceux aussi de SAS et Stata, savent qu'ils peuvent créer des procédures pour se faciliter la vie et ce, d'autant plus qu'ils auront des tâches répétitives à effectuer.

Ce parcours transversal entre informatique, méthode et statistique est riche d'enseignements. Il n'est pas propre aujourd'hui à la cancérologie, mais concerne tous les domaines de la médecine et de la recherche. Dans la pratique des grandes unités de recherche clinique, les « briques » informatiques sont définies de manière très formalisée, dans un souci d'efficacité (rapport coûts/fonctionnalités). Le morcellement des fonctions impose de lui-même les solutions logicielles tandis que la préservation des compétences interdit les changements intempestifs de ces dernières. L'utilisation d'outils standards est une des solutions pour faciliter l'organisation du service par des procédures qui permettent de garder une traçabilité des actions. Cette solution s'avère importante en cas de renouvellement des personnels. Sachant que pour un grand laboratoire pharmaceutique, le coût des investissements logiciels est relativement marginal par rapport à ceux de développement d'une nouvelle molécule, les choix se portent fréquemment sur des solutions informatiques auxquelles ne peuvent prétendre les organismes de santé académiques.

Dans les structures de moindre importance, une négociation est possible quant aux solutions envisagées. Mais la recherche clinique demeure un domaine où la rigueur méthodologique, la gestion des risques est de mise et où l'on ne peut garantir un résultat scientifique si l'environnement informatique s'avère inconsistant, voire incohérent. Dès lors qu'un établissement de santé est promoteur d'un essai, soit il met en place une unité dotée de moyens informatiques et personnels suffisants, soit il mutualise avec d'autres la gestion de sa recherche, à moins qu'il ne l'externalise. À l'intérieur de ces limites, diverses solutions logicielles peuvent être envisagées et le clinicien et/ou le statisticien disposent d'une marge de manœuvre.

Pour les recherches moins lourdes, c'est-à-dire qui ne passent pas par la mise en place d'un essai clinique prospectif, l'éventail des approches informatiques est presque sans limite. Mais la statistique reste une spécialisation et le chercheur, quelles que soient ses qualités, ne pourra, à notre avis, échapper soit à une formation méthodologique conséquente, soit au recours à un statisticien de formation. Sans doute est-ce à ce niveau que l'on peut utiliser le plus les ressources offertes gratuitement sur Internet, charge au chercheur de se constituer rapidement une bibliothèque d'outils qu'il n'aura plus besoin de (re)découvrir à chaque utilisation. Ce type de solution trouve néanmoins une limite peu prévisible au départ : l'interfaçage des outils sélectionnés peut être rapidement chronophage et s'avérer un obstacle incompatible avec l'économie recherchée.

Le développement de programmes statistiques nous semble devoir rester limité aux chercheurs en statistique ou en informatique, principalement quand des besoins très ciblés doivent être satisfaits. C'est le cas quand le marché n'offre pas de solution correspondant au problème à résoudre (modélisation, etc.) ou quand les logiciels existants sont très onéreux. Nous avons parlé précédemment de statistiques relatives à la généalogie dont s'occupe l'oncogénétique. Mais d'autres domaines de pointe peuvent nécessiter des développements logiciels, en particulier quand on met au point une nouvelle approche. Souvent ces travaux peuvent faire l'objet d'un travail de thèse pendant lequel le doctorant acquiert une compétence mixte, aussi bien statistique qu'informatique.

Ce parcours montre aussi le nombre de développements informatiques réalisés par la communauté scientifique mondiale, et l'on ne peut que saluer l'effort des universités, des établissements de santé publics, et parfois de simples développeurs, pour mettre gracieusement à disposition du monde entier le fruit de leur labeur et ce, en dépit de la concurrence entre les centres de recherches.

Références

1. Vickers AJ. A basic introduction to research: How not to do research. *J Soc Integr Oncol* 2008 ; 6 (2) : 82-5.

2. Le Tourneau C, Jack Lee J, Siu LL. Dose escalation methods in Phase I cancer clinical trials. *J Natl Cancer Inst* 2009 ; 101 : 708-20.

3. Rogatko A, Schoeneck D, Jonas W, Tighiouart M, Khuri FR, Porter A. Translation of innovative designs into phase I trials. *J Clin Oncol* 2007 ; 25 (31) : 4982-6.

4. O'Quigley J, Pepe M, Fisher L. Continual reassessment method: A practical design for phase I in cancer. *Biometrics* 1990 ; 46 : 561-6.

5. O'Quigley J, Shen LZ. Continual reassessment method: A likelihood approach. *Biometrics* 1996 ; 52 (2) : 673-84.

6. Kramar A, Houédé N, Paoletti X. np1: A computer program for dose escalation strategies in phase I clinical trials. *Comput Methods Programs Biomed* 2007 ; 88 (1) : 8-17.

7. Babb J, Rogatko A, Zacks S. Cancer phase I clinical trials: Efficient dose escalation with overdose control. *Stat Med* 1998 ; 17 (10) : 1103-20.

8. Simon R. Optimal two stage designs for phase II clinical trials. *Control Clin Trials* 1989 ; 10 : 1-10.

9. Bryant J. Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995, 51 (4) : 1372-83.

10. Fleming T. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982 ; 38 : 143-52.

11. Ensign L, Gehan E, Kamen D, Thall P. An optimal three-stage design for phase II clinical trials. *Stat Med* 1994 ; 13 : 1727-36.

12. Machin D, Campbell MJ, Tan SB, Tan SH. *Sample size tables for clinical studies*. 3rd edition. Oxford : Wiley-Blackwell, 2009.

13. Deligdisch L. Effects of hormones therapy on the endometrium. *Mod Pathol* 1993 ; 6 (1) : 94-106.

14. Kwiatkowski F, Slim K, Verrelle P, Chamorey E, Kramar A. Le score de propension : intérêt et limites. *Bull Cancer* 2007 ; 94 : 680-6.

15. Schwartz D, Flamant R, Lellouch J. *L'essai thérapeutique chez l'homme*. 2e édition. Paris : Flammarion Médecine-Sciences, 1981 : 93-128.

16. Patel S, Lyons-Weller J. CaGEDA : A web application for the integrated analysis of global gene expression patterns in cancer. *Applied Bioinformatics* 2004 ; 3 (1) : 49-62.

17. Saeed Al, Sharov V, White J, *et al*. TM4 : A free open-source system for microarray data management and analysis. *Biotechniques* 2003 ; 34 (2) : 374-8.

18. Darmoni SJ, Ladner J, Devos P, Gehanno JF. Robustesse du score SIGAPS, critère de bibliométrie pour valoriser les publications des établissements de santé. *Presse Med* 2009 ; 38 : 1058-61.